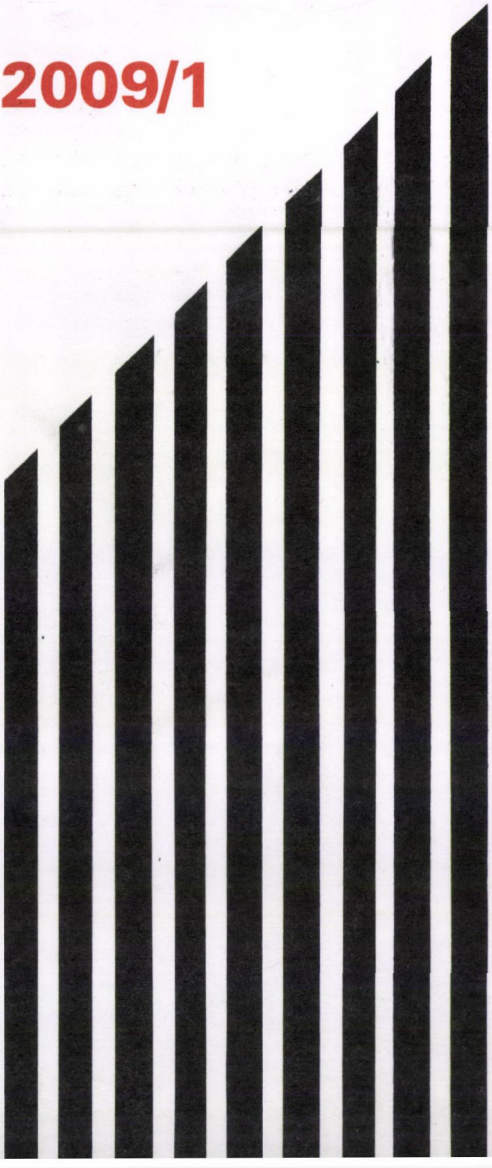
Alkalmazott matematikai lapok

2009/1

A MAGYAR TUDOMÁNYOS AKADÉMIA
MATEMATIKAI TUDOMÁNYOK
OSZTÁLYÁNAK KÖZLEMÉNYEI



26.

KÖTET

ALKALMAZOTT MATEMATIKAI LAPOK

A MAGYAR TUDOMÁNYOS AKADÉMIA MATEMATIKAI TUDOMÁNYOK OSZTÁLYÁNAK KÖZLEMÉNYEI

ALAPÍTOTTÁK
KALMÁR LÁSZLÓ, TANDORI KÁROLY, PRÉKOPA ANDRÁS, ARATÓ MÁTYÁS

FŐSZERKESZTŐ
PÁLES ZSOLT

FŐSZERKESZTŐ-HELYETTESEK
BENCZÚR ANDRÁS, SZÁNTAI TAMÁS

FELELŐS SZERKESZTŐ
VIZVÁRI BÉLA

TECHNIKAI SZERKESZTŐ
KOVÁCS GERGELY

26. kötet

Szerkesztőség és kiadóhivatal: 1027 Budapest, Fő u. 68.

Az Alkalmazott Matematikai Lapok változó terjedelmű füzetekben jelenik meg, és olyan eredeti tudományos cikkeket publikál, amelyek a gyakorlatban, vagy más tudományokban közvetlenül felhasználható új matematikai eredményt tartalmaznak, illetve már ismert, de színvonalas matematikai apparátus újszerű és jelentős alkalmazását mutatják be. A folyóirat közöl cikk formájában megírt, új tudományos eredménynek számító programokat, és olyan, külföldi folyóiratban már publikált dolgozatokat, amelyek magyar nyelven történő megjelentetése elősegítheti az elért eredmények minél előbbi, széles körű hazai felhasználását. A szerkesztőbizottság bizonyos időnként lehetővé kívánja tenni, hogy a legjobb cikkek nemzetközi folyóiratok különszámaként angol nyelven is megjelenhessenek.

A folyóirat feladata a Magyar Tudományos Akadémia III. (Matematikai) Osztályának munkájára vonatkozó közlemények, könyvismertetések stb. publikálása is.

A kéziratok a főszerkesztőhöz, vagy a szerkesztőbizottság bármely tagjához beküldhetők. A főszerkesztő címe:

Páles Zsolt, főszerkesztő
1027 Budapest, Fő u. 68.
A folyóirat e-mail címe: aml@math.elte.hu

Közlésre el nem fogadott kéziratokat a szerkesztőség lehetőleg visszajuttat a szerzőhöz, de a beküldött kéziratok megőrzéséért vagy továbbításáért felelősséget nem vállal.

Az Alkalmazott Matematikai Lapok előfizetési ára évfolyamonként 1200 forint. Megrendelések a szerkesztőség címén lehetségesek.

A Magyar Tudományos Akadémia III. (Matematikai) Osztálya a következő idegen nyelvű folyóiratokat adja ki:

1. Acta Mathematica Hungarica,
2. Studia Scientiarum Mathematicarum Hungarica.

KÜLÖNSZÁM

A kötetben megjelenő cikkek Az áramlástan numerikus módszerei: elmélet és alkalmazások című konferencián hangzottak el (2006. november, Győr, Széchenyi István Egyetem).

Az Alkalmazott Matematikai Lapok szerkesztőbizottsága köszönetet mond Horváth Zoltánnak (SZE, Győr), aki a kötet elkészültét vendég szerkesztőként segítette.

Alkalmazott Matematikai Lapok **26** (2009), 151-177.

VÉGESELEM MÓDSZER POLIMER-FOLYADÉKOK DETERMINISZTIKUS MODELLJÉHEZ

DAVID KNEZEVIC, SÜLI ENDRE[1]

E cikkben megvizsgálunk egy oldott polimer-folyadékok dinamikáját leíró, többskálájú, csatolt Navier–Stokes (NS) és Fokker–Planck (FP) modell megoldására javasolt végeselem alapú módszert. A szakirodalomban nem sokan foglalkoztak a probléma mienkhez hasonló, determinisztikus megközelítésével, mivel ez hihetetlen számítási kihívást jelent amiatt, hogy a Fokker–Planck-egyenlet analitikus megoldása egy nagyon sok változótól függő leképezés is lehet. Például, háromdimenziós áramlás olyan szimulációja esetén, amely polimer molekulák súlyzó modelljén alapul, egy olyan csatolt NS–FP-rendszert kell megoldanunk, ahol a Fokker–Planck-egyenletet hatdimenziós tartományon állítjuk fel.

A cikkben először áttekintjük az NS–FP-modell fizikai és matematikai alapjait, majd részletesen kidolgozzuk az általunk javasolt determinisztikus végeselem alapú megoldási módszert. Bemutatunk néhány parallel számítással elért numerikus eredményt azért, hogy módszerünk hatékonyságát demonstráljuk.

1. Bevezetés

A polimer-folyadékok dinamikája fejlett kutatási terület, melyet több mint hetven éve tanulmányoznak. Az érdeklődést részben e folyadékok ipari és keresdelmi alkalmazásai sarkallják, de emellett a polimer-folyadékok dinamikája egyre nagyobb figyelmet kap alkalmazott matematikával és numerikus analízissel foglalkozóktól e terület adta különleges kihívásoknak köszönhetően. Ugyanis a polimer-folyadékok dinamikája alapvetően „többskálájú", hiszen ahhoz, hogy hűen modellezzük e folyadékok produkálta bonyolult rheológiai tulajdonságokat, kombinálnunk kell a mikroszkopikus polimer molekulák dinamikáját a folyadék egészének áramlásával. Cikkünk elsődleges célja az, hogy megpróbálja kiterjeszteni a többskálájú modellek matematikai és számítási módszereit polimer-folyadékok esetében. A polimer-folyadékok természetes belső bonyolultsága, valamint a kapcsolodó matematikai modellek többségének ennek eredményeképpen analitikusan nehezen kezelhető volta miatt a numerikus megközelítések egyre nagyobb szerepet játszanak.

[1]A szerkesztők köszönetüket fejezik ki Szilárd Ágnesnek (Rényi Alfréd Matematikai Kutatóintézet) és Gáspár Csabának (Széchenyi István Egyetem) az eredetileg angol nyelvű kézirat gondos, szakértő fordításáért.

Cikkünkben végeselem módszert dolgozunk ki oldott polimer-folyadékok numerikus modelljeire. A módszer tárgyalását azonban a 2. részre hagyjuk, mivel először rövid áttekintést adunk a polimer-folyadékok dinamikáját leíró elméletről. Determinisztikus megközelítési módszerünkből eredő számításainkat a 3. részben, míg következtetéseinket és további terveinket a 4. részben tárgyaljuk.

1.1. Polimer-folyadékok

A polimer molekulák – melyekre gyakran makromolekulaként hivatkoznak – alap-szerkezeti egységek, ún. monomerek, ismétlődő láncából állnak. E makromolekulák tulajdonságai okozzák, hogy a polimer-folyadékok nagyon eltérően viselkednek a newtoni folyadékokhoz képest. Viszko-elasztikus folyadékoknak hívják ezeket, evvel is kiemelve, hogy mind viszkózus, mind elasztikus tulajdonságokkal rendelkeznek. (Elasztikusak olyan értelemben, hogy ezek a folyadékok „emlékeznek" korábbi deformációikra.) Ez a viszko-elaszticitás eredményez olyan egzotikus jelenségeket, mint a nyíróvékonyodás (shear-thinning), a rúdra mászás (rod-climbing) és a csőnélküli szifon (tubeless syphon).

E cikkben oldott polimer-folyadékokkal foglakozunk, azaz feltételezzük, hogy a polimer molekulák egy newtoni közegben úsznak olyan alacsony koncentrációban, hogy az egyes polimer molekulákról feltehető, hogy nem érintkeznek egymással. Az ilyen folyadékok esetében a megmaradási egyenletek ugyanazok, mint a newtoni esetben, azaz a tömegmegmaradásra azt kapjuk, hogy

$$\underset{\sim}{\nabla} \cdot \underset{\sim}{u} = 0, \tag{1}$$

a momentum megmaradásra pedig

$$\rho \left(\frac{\partial \underset{\sim}{u}}{\partial t} + \underset{\sim}{u} \cdot \underset{\sim}{\nabla} \underset{\sim}{u} \right) = \underset{\sim}{\nabla} \cdot \underset{\approx}{\sigma}. \tag{2}$$

A rugalmassági alaptörvény

$$\underset{\approx}{\sigma} = -p \underset{\approx}{I} + \eta_s (\underset{\sim}{\nabla} \underset{\sim}{u} + (\underset{\sim}{\nabla} \underset{\sim}{u})^{\mathrm{T}}) + \underset{\approx}{\tau} \tag{3}$$

a folyadékon belüli feszültségi és deformációs tenzorok közötti összefüggés newtoni feltételeiből adódik egy plusz tag hozzáadásával. Ez a $\underset{\approx}{\tau}$ tag, a polimer extrafeszültsége, mely a polimer molekulák jelenlétének köszönhető.

Tegyük fel, hogy a folyadék az Ω fizikai tartományon belül marad, melyről feltesszük, hogy $\mathbb{R}^d$ egy korlátos, nyílt részhalmaza, $d = 2, 3$. Tegyük fel továbbá, hogy adottak valamilyen megfelelő peremfeltételek $\partial\Omega$-án; például tapadó fal feltételt teszünk fel vagy periodikus peremfeltételt. Akkor az (1–3) egyenletek a Navier–Stokes-egyenletekre vezetnek, ahol $\underset{\approx}{\tau}$ a forrás tag, s így az a feladat, hogy találjunk $\underset{\sim}{u} : (\underset{\sim}{x}, t) \in \Omega \times \mathbb{R} \to \underset{\sim}{u}(\underset{\sim}{x}, t) \in \mathbb{R}^d$ és $p : (\underset{\sim}{x}, t) \in \Omega \times \mathbb{R} \to p(\underset{\sim}{x}, t) \in \mathbb{R}$

leképezéseket, amikre

$$\rho\left(\frac{\partial \underset{\sim}{u}}{\partial t} + \underset{\sim}{u}\cdot\nabla\underset{\sim}{u}\right) - \eta_s\Delta\underset{\sim}{u} + \nabla p = \nabla\cdot\underset{\approx}{\tau} \qquad \Omega\times(0,T]-\text{ben}, \tag{4}$$

$$\nabla\cdot\underset{\sim}{u} = 0 \qquad \Omega\times(0,T]-\text{ben}, \tag{5}$$

$$\underset{\sim}{u}(\underset{\sim}{x},0) = \underset{\sim}{u}_0(\underset{\sim}{x}) \qquad \forall \underset{\sim}{x}\in\Omega.$$

A fenti egyenletben η_s a közeg viszkozitási együtthatója.

Ahhoz, hogy megoldjuk ezt az egyenletrendszert, meg kell határoznunk $\underset{\approx}{\tau}$-t. A hagyományos megközelítésmód egy olyan alaptörvény (általában algebrai vagy differenciálegyenlet) felállítása, mely csak makroszkopikus mennyiségektől függ, s amely összefüggést ad $\underset{\approx}{\tau}$ és a folyadék deformációs előtörténete között [?]. Egy ilyen alapegyenlet vagy alaptörvény alapulhat tisztán makroszkopikus megfontolásokon is, de manapság gyakoribb, hogy kinetikus elméletből vezetik le, mivel a kinetikus elméleten alapuló elemzés nagyobb modellezési szabadságot biztosít és bizonyítottan valósághűbb modellekhez vezet. Azonban – a legegyszerűbb eseteket kivéve, mint amilyen a Hooke-féle súlyzó modell (ld. az 1.2. részt) – ahhoz, hogy kinetikus elméleten alapuló modellből tisztán makroszkopikus alaptörvényt állítsunk fel, a modell közelítő lezártjával kell dolgoznunk, ez pedig az alapmodell pontosságát rontja [?].

Az, hogy hagyományosan a kutatások az alapmodell közelítő lezártjára fókuszáltak, érthető, mivel az olyan modellek megoldásainak kiszámítása, melyek csak makroszkopikus változóktól függenek, sokkal kevesebb munkát igényel, s számos esetben (főleg egyszerű áramlások esetén) e makroszkopikus modellek analitikusan is megoldhatók. Azonban a rendelkezésre álló számítási lehetőségek robbanásszerű megnövekedése miatt lehetővé és kívánatossá vált, hogy közvetlenül a pontosabb, többskálájú, a kinetikus elméletet a mikroszkopikus szinttel összekötő „mikro-makro" modellek alapján számítsunk. E cikk középpontjában a többskálájú modellek állnak; ezért a továbbiakban nem foglalkozunk tisztán makroszkopikus modellekkel. Le szeretnénk szögezni azonban, hogy a makroszkopikus megközelítésmód továbbra is alapvető része a folyadékok elméleti és numerikus rheológiai kutatásának, s a kapcsolódó számítások hatékonysága, illetve a makroszkopikus modellek matematikai kezelhetősége biztosítják, hogy e modellek a belátható jövőben továbbra is fontos szerepet játszanak.

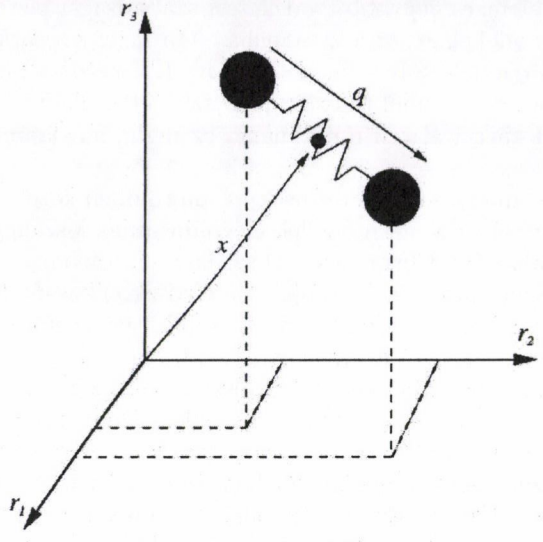
1.2. Polimer-láncok modelljei

Ahhoz, hogy kinetikus elméleten alapuló egyenleteket állíthassunk fel a polimer-folyadékok viselkedésének leírására, arra van szükség, hogy az egyes polimer molekulákat egyszerű modell írja le. Az elmúlt ötven évben polimerek számos „durva-szemcsés" mechanikai modelljét vezették be (a legfontosabb példák tárgyalását lásd Bird és munkatársai [?] 10. fejezetében). A két leginkább használatos az ún. *Kramer-lánc* [?], illetve a *Rouse–Zimm-lánc* [?, ?] modell. Mindkét modell

golyók láncaként ábrázolja a polimert. A Kramer-lánc esetében a golyókat merev, tömeggel nem rendelkező rudak kötik össze, míg a Rouse–Zimm-lánc esetében ($\underset{\sim}{F}$ kifejtett erővel rendelkező) rugókat használnak a golyók összekapcsolására.

Ezen modellek numerikus szimulációja azonban nagyon drága, mivel általában nagy szabadságfokkal rendelkeznek (körülbelül tíztől többszázig). Ennek eredményeképpen a legnagyobb figyelmet a polimer modellek hierarchiájának legalján álló, legegyszerűbb ún. *súlyzó* modell kapta. Ez a modell csak két golyóból áll, melyeket egy rugó köt össze. A súlyzót teljesen meghatározza a tömegközéppont $\underset{\sim}{x}$ helye, valamint a végpontjai közötti $\underset{\sim}{q}$ vektor. Az $\underset{\sim}{x}$ lehetséges értékeit (azaz értelmezési tartományát) *fizikai térnek* nevezzük (jele Ω), míg $\underset{\sim}{q}$-ét *konfigurációs térnek* (jele D).

Az 1. ábrán látható egy súlyzó sematikus képe, s az $\underset{\sim}{x}$ és $\underset{\sim}{q}$ változókat is bejelöltük.



1. ábra. A súlyzó modell két golyóból és egy összekötő rugóból áll. A súlyzó állapotát meghatározza a tömegközéppont $\underset{\sim}{x}$ helyzete és a végpontokat összekötő $\underset{\sim}{q}$ vektor.

A súlyzó modell egyszerűsége ellenére is nagyon hasznos a polimer-folyadékok viselkedésének leírásakor. Ez azért van, mert a súlyzót nyújthatja és irányíthatja az áramlás, s e két tulajdonság sok esetben nagyjából meghatározza a polimer-folyadék rheológiai tulajdonságait.

A súlyzó modell függ az összekötő rugót leíró erőtörvénytől. A két legáltalánosabban használatos a „Hooke-rugó", ahol

$$\hat{\underset{\sim}{F}}(\underset{\sim}{q}) = \underset{\sim}{q},$$

illetve a *végesen nyújtható, nemlineáris, elasztikus* (Finitely Extensible Nonlinear Elastic (FENE)) rugó [?], melyre

$$\underset{\sim}{F}(\underset{\sim}{q}) = \frac{\underset{\sim}{q}}{1 - |\underset{\sim}{q}|^2/b}. \tag{6}$$

Fentebb mindkét egyenlet dimenzió nélküli formáját írtuk fel.

Hooke-rugók esetében $\underset{\sim}{q} \in \mathbb{R}^d$, míg FENE-rugóknál $\underset{\sim}{q} \in B(0, \sqrt{b})$, ahol $B(0, s)$ a d-dimenziós, origó középpontú, s sugarú gömböt jelöli, b pedig egy dimenzió nélküli paraméter, melynek értékkészlete általában $[10, 1000]$ (ld. [?]), s amely a rugó maximális megnyúlását határozza meg.

A Hooke-rugók esetét széleskörűen és alaposan tanulmányozták, mivel analitikusan megoldható egyenletekhez vezet. Kiemeljük például, hogy a jólismert Oldroyd-B-modell, mely oldott polimer-folyadékok leírására szolgál, s melyet eredetileg kontinuum-mechanikai meggondolások alapján állítottak fel [?], ekvivalens a Hooke-féle mikro–makro súlyzó modellel [?]. Mivel azonban a Hooke-féle lineáris törvénynek elegettevő rugók a valóságban nem nyújthatók a végtelenségig, ez a modell bizonyos esetekben, mint amilyen az erősen nyújtó hatású áramlás, nem használható.

Ilyen esetben használhatjuk a FENE-modellt, melyben korlátozott a nyújthatóság. A FENE-modell valósághűbb, mint a Hooke-modell, de a (nemlineáris) FENE-modell használatakor fel kell adnunk minden reményt arra, hogy zárt alakú analitikus eredményt kapjunk nem-egyensúlyi áramlás esetében. Így FENE-rugóknál numerikus megközelítésre van szükség.

Célunk egy hatékony végeselem módszeren alapuló numerikus eljárás kidolgozása és elemzése, melynek segítségével megoldható az oldott polimer-folyadékok olyan többskálájú modelljéből adódó egyenletrendszer, ahol a polimereket FENE-rugóval összekötött súlyzóknak képzeljük. Ez jelentős kihívást jelent még úgy is, hogy a súlyzó modell a polimer molekuláknak egy elég durva megközelítése. Például, háromdimenziós áramlás esetén olyan egyenletrendszert kell megoldanunk, ahol a Navier–Stokes és Fokker–Planck-egyenletek csatolt rendszert alkotnak, s ahol a Fokker–Planck-egyenlet hatdimenziós, mivel az egyes súlyzók szabadságfoka éppen hat: 3 az $\underset{\sim}{x}$ és 3 a q változóból adódóan. Amennyiben az általunk kidolgozott módszer sikeresen alkalmazható a súlyzó modellből következő egyenletekre, akkor további kutatás célja lehet a módszer kiterjesztése golyó-rugó típusú láncok esetére.

1.3. A Fokker–Planck-egyenlet

A polimermolekula-konfiguráció ψ valószínűségi sűrűségfüggvényének evolúcióját Fokker–Planck-egyenlet írja le. Az egyenlet részletesen kidolgozott levezetése

megtalálható Lozinski [?] Ph.D. dolgozatában és Barrett & Süli [?] cikkében. Most csak a végeredményt idézzünk. Tehát keressük azt a

$$\psi : (\underset{\sim}{x}, \underset{\sim}{q}, t) \in \Omega \times D \times (0,T] \to \psi(\underset{\sim}{x}, \underset{\sim}{q}, t) \in \mathbb{R}_{\geq 0}$$

leképezést, amire

$$\frac{\partial \psi}{\partial t} + (\underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x)\psi + \underset{\sim}{\nabla}_q \cdot \left(\left(\underset{\approx}{\kappa} \underset{\sim}{q} - \frac{1}{2\lambda} \underset{\sim}{F}(\underset{\sim}{q}) \right) \psi \right) = \frac{1}{2\lambda} \Delta_q \psi. \tag{7}$$

A fenti egyenletben $\underset{\sim}{F}$ jelöli a súlyzó összekötő rugója erejét, λ pedig a polimerre jellemző relaxációs időt. Minden $(\underset{\sim}{x}, t) \in \Omega \times [0,T]$ pont esetén $\psi(\underset{\sim}{x}, \cdot, t)$ egy valószínűségi sűrűségfüggvény, ezért kielégíti az alábbi normalizációs feltételt:

$$\int_D \psi(\underset{\sim}{x}, \underset{\sim}{q}, t) \, \mathrm{d}\underset{\sim}{q} = 1.$$

A (7) egyenletet kiegészítjük a $\psi(\underset{\sim}{x}, \underset{\sim}{q}, 0) = \psi_0(\underset{\sim}{x}, \underset{\sim}{q})$ kezdeti feltétellel és megfelelő peremfeltételekkel. A (7) egyenlet viselkedése nagyon különböző a fizikai és a konfigurációs térben, s a peremfeltételeknek ezt tükrözniük kell. Ugyanis a konfigurációs térben a (7) egyenlet parabolikus és általában homogén Dirichlet-peremfeltételt adunk meg $\Omega \times \partial D$-n, mivel a súlyzó nem éri el a maximális lehetséges hosszát.

A fizikai térben azonban hiperbolikus egyenletet kapunk, s ezért, ha $\partial\Omega^-$ jelöli a perem azon részét, ahol a folyadék beáramlik, akkor csak $\partial\Omega^- \times D$-n adunk meg peremfeltételt. Ahhoz azonban, hogy csak a perem azon részén adjunk peremfeltételt, ahol a folyadék beáramlik, ismernünk kell ψ-t a peremen, folyásiránnyal szemben. Egy lehetséges megoldás, hogy periodikus határral dolgozunk. Egy másik megoldás, hogy teljesen kifejlődött áramlást tételezünk fel a folyásiránnyal szemben azért, hogy az áramlás valószínűségi sűrűségfüggvényét meg tudjuk határozni.

1.4. A polimer extra-feszültségi tenzora

A (7) egyenlet bevezetésének célja az volt, hogy segítségével ki tudjuk számítani a $\underset{\approx}{\tau}(\underset{\sim}{x}, t)$ extra-feszültségi tenzort. Biller és Petruccione [?] kidolgozták azt a képletet (az ún. *Kramers-képlet* általánosítását), mellyel $\underset{\approx}{\tau}$ kiszámítható ψ-ből inhomogén sebességmező esetén:

$$\underset{\approx}{\tau}(\underset{\sim}{x}, t) = n_p k T \left(-\underset{\approx}{I} + \int_D \underset{\sim}{q} \otimes \underset{\sim}{F}(\underset{\sim}{q}) \psi(\underset{\sim}{x}, \underset{\sim}{q}, t) \, \mathrm{d}\underset{\sim}{q} \right), \tag{8}$$

ahol n_p az ún. polimerszám sűrűség, k a Boltzmann-állandó, és T az abszolút hőmérséklet. A (8) egyenletből és a (6) kifejezésből látható, hogy $\underset{\approx}{\tau}$ szimmetrikus.

Célunk szempontjából hasznos a (8) képletet kifejezni az ún. *polimer viszkozitás* segítségével, melynek jele η_p, s melyet a newtoni folyadékok viszkozitásához hasonlóan definiálnak. FENE súlyzó modell esetében, $\dot{\gamma}$ nyírósebességű nyíróáramlást feltételezve, megmutatható, hogy a nyírásirányú feszültséget jól közelíti

$$\tau_{xy} \approx \dot{\gamma}\lambda n_p kT \left(\frac{b+d+2}{b}\right);$$

(ld. [?].) Így a polimer viszkozitás:

$$\eta_p := \lambda n_p kT \left(\frac{b+d+2}{b}\right).$$

A fentiek felhasználásával a (8) képlet a következőt adja FENE súlyzókra:

$$\underset{\approx}{\tau}(\underset{\sim}{x},t) = \frac{\eta_p}{\lambda}\left(\frac{b+d+2}{b}\right)\left(-\underset{\approx}{I} + \int_D \underset{\sim}{q}\otimes\underset{\sim}{F}(\underset{\sim}{q})\psi(\underset{\sim}{x},\underset{\sim}{q},t)\,\mathrm{d}\underset{\sim}{q}\right).$$

A csatolt NS–FP-egyenletrendszer időben globális, gyenge megoldásai létezésére vonatkozó analitikus eredményeket illetően lásd Barrett, Schwab & Süli [?] és Barrett & Süli [?] cikkeit, melyekben a vonatkozó elmélet fejlődésének részletes leírása is megtalálható.

Most, hogy áttekintettük a csatolt NS–FP-rendszert és meghatároztuk $\underset{\approx}{\tau}$-t, készen állunk e mikro-makro rendszer numerikus megoldásának tárgyalására.

2. Polimer-folyadékok dinamikájának numerikus megközelítései

A numerikus rheológia hőskora kb. 1970-re tehető. A korai kutatások szükségszerűen kizárólag az 1. részben tárgyalt makroszkopikus megközelítéssel dolgoztak, mivel ez a számítások szempontjából sokkal kevésbé idő- és eszközigényes, mint a mikro-makro módszerek. A makroszkopikus számítások tipikusan a folyadékok dinamikájának szokásos számítási eszközeit használják, mint amilyenek a végeselem, végestérfogat és spektrál módszerek. Az ilyen irányú kutatások jelentős mértékűek, s a terület még mindig aktívan fejlődik; lásd Keunings [?] informatív összefoglalóját.

Alternatív megközelítésként, a kora 1990-es évektől megnőtt a többskálájú modellek (mint amilyen az 1. részben tárgyalt csatolt NS- és FP-egyenletrendszer) közvetlen vizsgálatának népszerűsége. A kulcsötletet e téren Öttinger és Laso adták 1992-ben (ld. [?]), amikor azt javasolták, hogy használjuk ki a (7) Fokker–Planck-egyenlet ekvivalenciáját a

$$\mathrm{d}\underset{\sim}{q}(\underset{\sim}{x},t) + \underset{\sim}{u}(\underset{\sim}{x},t)\cdot\underset{\sim}{\nabla}_x\underset{\sim}{q}(\underset{\sim}{x},t)\mathrm{d}t = \\ = \left(\underset{\approx}{\kappa}(\underset{\sim}{x},t)\underset{\sim}{q}(\underset{\sim}{x},t) - \frac{1}{2\lambda}\underset{\sim}{F}(\underset{\sim}{q}(\underset{\sim}{x},t))\right)\mathrm{d}t + \sqrt{\frac{1}{\lambda}}\,\mathrm{d}\underset{\sim}{W}(\underset{\sim}{x},t) \quad (9)$$

Itô sztochasztikus differenciálegyenlettel, majd oldjuk meg a (9) egyenletet Monte-Carlo típusú módszerrel. Ez az ötlet számos módszert eredményezett, melyeket összefoglalóan *sztochasztikus módszereknek* nevezünk, s melyeket teljesen kidolgoztak és polimer-folyadékok széles skálájának modellezésére használtak (ld. például [?, ?, ?]). A sztochasztikus megközelítésnek van azonban egy igen gyenge pontja: a fellépő sztochasztikus hiba csak lassan csökken (tipikusan $\mathcal{O}\left(N^{-1/2}\right)$ nagyságrendben, ha $N \to \infty$, ahol N a módszerben használt pontok száma).

Ugyan kidolgoztak variáció csökkentő eljárásokat azért, hogy ezt a hibatagot javítsák, de a sztochasztikus hiba jelenléte még eme eljárások alkalmazása után is hátrányt jelent, s elkerülése fontos motivációs tényező a determinisztikus módszerekre való átállásra. Másrészt azonban a sztochasztikus megközelítés egy jelentős előnye, hogy jól illeszkedik a polimer-modell szabadságfokaihoz – például többszáz szabadságfokú modellekkel is végeztek számításokat (ld. [?]).

Az e cikkben propagált többskálájú megközelítésként szeretnénk *determinisztikusan* meghatározni mind az $\underset{\sim}{u}(\underset{\sim}{x}, t)$ sebességmezőt, mind a $\psi(\underset{\sim}{x}, \underset{\sim}{q}, t)$ sűrűségfüggvényt. A legfőbb nehézséget a Fokker–Planck-egyenlet nagy dimenziójának kezelése jelenti, mely súlyzóval modellezett szuszpenzió háromdimenziós áramlása esetén hatdimenziós. Természetesen a dimenzió növekszik, ha olyan mechanikai modellt alkalmazunk, melyben nő a szabadságfok. Viszonylag kevesen vizsgálták ezt a megközelítésmódot – nagy valószínűséggel éppen a nagy dimenziószám volt elrettentő hatással arra, hogy a FP-egyenletet közvetlenül próbálják megoldani. 1972-ben Stewart és Sørensen gömbi harmonikus polinomokat használtak a Fokker–Planck-egyenlet megoldására merev súlyzók oldott szuszpenziójának stacionáris nyíróáramlása esetén (ld. [?]). Warner hasonló módszert alkalmazott FENE-típusú súlyzók nyíróáramlásának tanulmányozására (ld. [?]), eredményeit pedig 13 évvel később Fan fejlesztette tovább (ld. [?]). E korai tanulmányok egyszerűsítésképpen csak homogén áramlást vettek figyelembe, amikoris ψ csak $\underset{\sim}{q}$ és t függvénye. Fan 1989-es cikke (ld. [?]) volt az első olyan munka, melyben determinisztikus megközelítéssel oldottak meg nem homogén $\underset{\sim}{u}$ vektormezőt. Ebben síkbeli csatorna áramlást szimulált golyó-rúd típusú polimer modellt használva. Amiatt, hogy Fan golyó-rúd típusú modellből indult ki számításai során, kétdimenziós konfigurációs térrel kellett dolgoznia. S bár ψ függött $\underset{\sim}{x}$-től, Fan egyszerűsítésképpen feltette, hogy $\underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \psi$, a fizikai tér konvektív tagja, eltűnik (azaz nulla). Fan eredményeit a későbbiekben kiterjesztették úgy, hogy már nem tették fel a fizikai tér konvektív tagjának eltűnését – Nayak a fizikai térben klasszikus Galerkin-módszert használt (ld. [?]), míg Grosso és munkatársai áramvonal-diffúziós módszert használtak a konvektív tag kezelésére (ld. [?]).

Lozinski és Chauvière munkatársaikkal 2003-tól kezdődően egy cikksorozatban jelentősen kiterjesztették a kurrens determinisztikus módszereket (ld. [?, ?, ?, ?, ?]). Egy a szerzők által alkalmazott fontos eljárás az, hogy a Fokker–Planck-egyenletet

minden időintervallumban két részre bontanak:

$$\frac{\tilde{\psi}-\psi^n}{\Delta t}+\underset{\sim}{\nabla}_q\cdot\left(\left(\underset{\approx}{\kappa}^n\underset{\sim}{q}-\frac{1}{2\lambda}\underset{\sim}{F}(q)\right)\tilde{\psi}\right)=\frac{1}{2\lambda}\Delta_q\tilde{\psi}, \tag{10}$$

$$\frac{\psi^{n+1}-\tilde{\psi}}{\Delta t}+\underset{\sim}{u}^n\cdot\underset{\sim}{\nabla}_x\psi^{n+1}=0. \tag{11}$$

A fenti kifejezésekben $\tilde{\psi}$ egy közbenső érték, $\underset{\sim}{u}^n$ és $\underset{\approx}{\kappa}^n=(\underset{\sim}{\nabla}_x\underset{\sim}{u}^n)$ pedig az n időpillanatban van kiértékelve.

Mi is alkalmazzuk ezt az operátor felbontási eljárást a Fokker–Planck-egyenlet végeselem alapú megoldása során (ld. a 2.1.2. részt). Háromdimenziós áramlás esetén például az eljárás azt eredményezi, hogy egy sor háromdimenziós feladatot kell megoldanunk a „teljes" hatdimenzós Fokker–Planck-egyenlet helyett. A dimenzió ilyetén való csökkentése révén mérsékeljük az ún. dimenziós átok hatását, mely a rácspontok számának (s ezáltal a probléma számítási bonyolultságának) a dimenzió növelésekor bekövetkező exponenciális növekedésére utal. Lozinski és Chauvière [?, ?, ?] cikkeikben megmutatták, hogy FENE súlyzó modell esetén az általuk javasolt determinisztikus eljárás hatékonyabb a sztochasztikus megközelítésnél bizonyos, a szakirodalomban használt tesztfeladat esetén. Például síkbeli csatornán belüli áramlást vizsgáltak köralakú akadállyal (ld. ??. részt, ahol a problémát tárgyaljuk). Egy sztochasztikus eljárást hasonlítottak össze determinisztikus megoldásukkal és megmutatták, hogy a determinisztikus megközelítés jelentősen hatékonyabb volt a számítási költségek szempontjából, valamint pontosabb is, a sztochasztikus hiba hiányának köszönhetően.

Lozinski és munkatársai eredményei bebizonyították, hogy alacsony dimenziós konfigurációs terű modellek esetén a determinisztikus megközelítés esetenként jobban „teljesít" a sztochasztikusnál. Az azonban még mindig nyitott kérdés, hogy háromnál nagyobb dimenziójú konfigurációs tér esetén a determinisztikus megközelítés mennyire hatékony.

2.1. Determinisztikus algoritmus mikro-makro modell esetére

Ebben a részben ismertetjük a csatolt Fokker–Planck és Navier–Stokes mikro-makro rendszer megoldására kifejlesztett numerikus módszerünket. Megjegyezzük, hogy Barrett, Schwab és Süli, illetve Barrett és Süli cikkeikben (ld. [?], [?]) számos eredményt értek el e rendszer gyenge megoldása létezésével kapcsolatban.

2.1.1. A Navier–Stokes-rendszer numerikus megközelítése

Tekintsük a Navier–Stokes-egyenleteket. Az egyszerűség kedvéért legyen a (4) egyenletben $\rho=1$, és követve [?]-t a (4) és (5) egyenletek gyenge megfogalmazása

szerint keressük (például homogén Dirichlet-peremfeltétel esetében) a

$$\underset{\sim}{u} \in [\underset{\sim}{V}]^d := \left[\mathrm{H}_0^1(\Omega)\right]^d \quad \text{és} \quad p \in \Pi = \left\{ q \in \mathrm{L}^2(\Omega) : \int_\Omega q \,\mathrm{d}\underset{\sim}{x} = 0 \right\}$$

függvényeket, amikre:

$$\int_\Omega \frac{\partial \underset{\sim}{u}}{\partial t} \cdot \underset{\sim}{v} \,\mathrm{d}\underset{\sim}{x} + \eta_s \int_\Omega \nabla_{\underset{\sim}{x}} \underset{\sim}{u} : \nabla_{\underset{\sim}{x}} \underset{\sim}{v} \,\mathrm{d}\underset{\sim}{x} - \int_\Omega (\nabla_{\underset{\sim}{x}} \cdot \underset{\sim}{v}) p \,\mathrm{d}\underset{\sim}{x}$$
$$+ \int_\Omega (\underset{\sim}{u} \cdot \nabla_{\underset{\sim}{x}} \underset{\sim}{u}) \cdot \underset{\sim}{v} \,\mathrm{d}\underset{\sim}{x} + \int_\Omega \underset{\approx}{\tau} : \nabla_{\underset{\sim}{x}} \underset{\sim}{v} \,\mathrm{d}\underset{\sim}{x} = 0, \tag{12}$$

és

$$\int_\Omega (\nabla_{\underset{\sim}{x}} \cdot \underset{\sim}{u}) q \,\mathrm{d}\underset{\sim}{x} = 0, \tag{13}$$

minden $\underset{\sim}{v} \in [\underset{\sim}{V}]^d = \times_{i=1}^d V$ és $q \in \Pi$ esetén, ahol $\underset{\approx}{A} : \underset{\approx}{B} = \sum_{i,j=1}^d A_{ij} B_{ij}$ az $\underset{\approx}{A}$ és $\underset{\approx}{B}$ mátrixok skalárszorzata. Vegyes, esetleg nem homogén, Dirichlet–Neumann-peremfeltétel esetén a peremfeladat gyenge alakja hasonló, de a V és Π függvényterek definiciói kissé megváltoznak. Az 1. részben tárgyaltak szerint $\underset{\approx}{\tau}$, a polimer extra-feszültségi tenzora, kiszámítható a Fokker–Planck-egyenlet megoldásából. Ez az a kifejezés, mellyel a mikro és makro egyenletek csatolódnak.

A továbbiakban feltesszük, hogy $\underset{\approx}{\tau}$ egy adott forrás tag. Az FP-egyenlet numerikus megoldási módszerének részletei, illetve $\underset{\approx}{\tau}$ kiszámításának leírása a 2.1.2. részben található.

Ahhoz, hogy az NS-egyenletek gyenge formáját végeselem módszerrel implementáljuk, (12) és (13) mindegyikét felbontjuk három egyenletre, $\underset{\sim}{u} = (u_x, u_y, u_z)$ minden egyes komponensének megfelelően. Itt u_x az $\underset{\sim}{u}$ sebességmező x komponense. Ekkor u_x-re azt kapjuk, hogy meg kell találnunk az $u_x \in V$ és $p \in \Pi$ leképezéseket, amikre

$$\int_\Omega \frac{\partial u_x}{\partial t} v_x \,\mathrm{d}\underset{\sim}{x} + \eta_s \int_\Omega \nabla_{\underset{\sim}{x}} u_x \cdot \nabla_{\underset{\sim}{x}} v_x \,\mathrm{d}\underset{\sim}{x} - \int_\Omega p \frac{\partial v_x}{\partial x} \,\mathrm{d}\underset{\sim}{x}$$
$$+ \int_\Omega \left(\underset{\sim}{u} \cdot \nabla_{\underset{\sim}{x}} u_x\right) v_x \,\mathrm{d}\underset{\sim}{x} + \int_\Omega \left(\tau_{xx} \frac{\partial v_x}{\partial x} + \tau_{xy} \frac{\partial v_x}{\partial y} + \tau_{xz} \frac{\partial v_x}{\partial z} \right) \mathrm{d}\underset{\sim}{x} = 0, \tag{14}$$

és

$$\int_\Omega q \frac{\partial u_x}{\partial x} \,\mathrm{d}\underset{\sim}{x} = 0, \tag{15}$$

minden $v_x \in V$ és $q \in \Pi$ esetén. Természetesen (14) és (15) hasonló csatolt egyenletekre vezetnek u_y és u_z szerint.

Legyen $\{\phi_1, \phi_2, \ldots, \phi_N\}$ a V vektortér egy V^h véges dimenziós alterének bázisa, s hasonlóan $\operatorname{span}\{\psi_1, \psi_2, \ldots, \psi_{N'}\} = \Pi^h$. A vegyes módszer stabilitásához (vagyis a Babuška–Brezzi-féle „inf-sup" feltétel teljesüléséhez, ld. [?]) a V^h alteret darabonkét kvadratikus, folytonos függvények alterének választjuk, míg Π^h darabonként lineáris, folytonos függvények altere. Térben és időben diszkretizálva az egyenleteket azt kapjuk, hogy $U_x^n(\underset{\sim}{x}) = \sum_j u_x^{n,j}\phi_j(\underset{\sim}{x})$ és $P^n(\underset{\sim}{x}) = \sum_j p^{n,j}\psi_j(\underset{\sim}{x})$. Itt az n felső index a $t = t^n = n\Delta t$ pillanatban vett értékre, a nagybetűs változók pedig a megfelő folytonos változók diszkrét verzióira utalnak. Az U_x^n és P^n változókat behelyettesítve a (14) és (15) egyenletekbe és időben retrográd Euler módszert használva azért, hogy az időintervallumok mérete tetszőleges maradjon stabilitás mellett, a diszkrét variációs probléma a következő: minden $n = 0, \ldots, M$ esetén (ahol $M = T/\Delta t$) találjunk olyan $\underset{\sim}{X}^n = (\underset{\sim}{u}_x^n, \underset{\sim}{u}_y^n, \underset{\sim}{u}_z^n, \underset{\sim}{p}^n)^{\mathrm{T}} \in \mathbb{R}^{3N+N'}$ vektort, amire

$$
\begin{aligned}
F_x^i(\underset{\sim}{X}^{n+1}) := {} & \sum_{j=1}^{N} u_x^{n+1,j} \int_{\Omega} \left(\phi_j\phi_i + \Delta t\eta_s \left(\nabla_{\underset{\sim}{x}}\phi_j \cdot \nabla_{\underset{\sim}{x}}\phi_i\right)\right) \mathrm{d}\underset{\sim}{x} \\
& + \Delta t \int_{\Omega} \left[\left(\sum_{j=1}^{N} u_x^{n+1,j}\phi_j\right) \left(\sum_{j=1}^{N} u_x^{n+1,j}\frac{\partial\phi_j}{\partial x}\right) \right. \\
& + \left(\sum_{j=1}^{N} u_y^{n+1,j}\phi_j\right) \left(\sum_{j=1}^{N} u_x^{n+1,j}\frac{\partial\phi_j}{\partial y}\right) \\
& \left. + \left(\sum_{j=1}^{N} u_z^{n+1,j}\phi_j\right) \left(\sum_{j=1}^{N} u_x^{n+1,j}\frac{\partial\phi_j}{\partial z}\right) \right] \phi_i \,\mathrm{d}\underset{\sim}{x} \\
& + \Delta t \int_{\Omega} \tau_{xx}\frac{\partial\phi_i}{\partial x} + \tau_{xy}\frac{\partial\phi_i}{\partial y} + \tau_{xz}\frac{\partial\phi_i}{\partial z} \,\mathrm{d}\underset{\sim}{x} \\
& - \Delta t \sum_{j=1}^{N'} p_j^{n+1} \int_{\Omega} \psi_j \frac{\partial\phi_i}{\partial x} \,\mathrm{d}\underset{\sim}{x} - \sum_{j=1}^{N} \int_{\Omega} u_x^{n,j}\phi_j\phi_i \,\mathrm{d}\underset{\sim}{x} = 0, \qquad (16)
\end{aligned}
$$

és

$$
G_x^{i'}(\underset{\sim}{X}^{n+1}) := \sum_{j=1}^{N} u_x^{n+1,j} \int_{\Omega} \psi_{i'} \frac{\partial\phi_j}{\partial x} \,\mathrm{d}\underset{\sim}{x} = 0, \qquad (17)
$$

minden $\phi_i, i = 1, \ldots, N$, és $\psi_{i'}, i' = 1, \ldots, N'$, esetén. Egyszerűsítésképpen $F_x^i(\underset{\sim}{X}^{n+1})$, valamint $G_x^{i'}(\underset{\sim}{X}^{n+1})$ jelöli a (16), illetve (17) egyenlet bal oldalát, s hasonlóképpen definiáljuk az $F_y^i(\underset{\sim}{X}^{n+1})$, $F_z^i(\underset{\sim}{X}^{n+1})$ és $G_y^{i'}(\underset{\sim}{X}^{n+1})$, $G_z^{i'}(\underset{\sim}{X}^{n+1})$ kifejezéseket u_y, illetve u_z esetén. A (16) és (17) egyenletek, illetve az y és z változó szerinti megfelelőik mindegyike egy vektort határoz meg, például $\underset{\sim}{F}_x(\underset{\sim}{X}) = (F_x^1(\underset{\sim}{X}), \ldots, F_x^N(\underset{\sim}{X}))^{\mathrm{T}} \in \mathbb{R}^N$. A $\underset{\sim}{G}_x, \underset{\sim}{G}_y, \underset{\sim}{G}_z \in \mathbb{R}^{N'}$ vektorokat egyetlen

$\underset{\sim}{G} \in \mathbb{R}^{N'}$ vektorba egyesítjük, melyet

$$\underset{\sim}{G}^{i'}(\underset{\sim}{X}^{n+1}) := \sum_{j=1}^{N} \left(u_x^{n+1,j} \int_\Omega \psi_{i'} \frac{\partial \phi_j}{\partial x} \, \mathrm{d}\underset{\sim}{x} + u_y^{n+1,j} \int_\Omega \psi_{i'} \frac{\partial \phi_j}{\partial y} \, \mathrm{d}\underset{\sim}{x} \right.$$
$$\left. + u_z^{n+1,j} \int_\Omega \psi_{i'} \frac{\partial \phi_j}{\partial z} \, \mathrm{d}\underset{\sim}{x} \right) = 0$$

határoz meg.

Legyen $\underset{\sim}{H} = (\underset{\sim}{F}_x, \underset{\sim}{F}_y, \underset{\sim}{F}_z, \underset{\sim}{G})^{\mathrm{T}} \in \mathbb{R}^{3N+N'}$. Ahhoz, hogy kiszámíthassuk a diszkrét NS-rendszer megoldását az $n+1$. pillanatban, a $\underset{\sim}{H}(\underset{\sim}{X}^{n+1}) = \underset{\sim}{0}$ nemlineáris egyenletrendszert kell megoldanunk. Ehhez Newton-módszert használunk. Jelölje J a rendszer Jacobi-determinánsát. A J mátrix elemeit $\underset{\sim}{H}(\underset{\sim}{X}^{n+1})$ vektor $\underset{\sim}{X}^{n+1}$ komponensei szerinti deriváltjai segítségével számítjuk ki. Tegyük fel, hogy az $\underset{\sim}{X}^{n+1}$ megoldás vektorunkat úgy rendezzünk, hogy az első N komponense megegyezik az $\underset{\sim}{u}_x^{n+1}$ vektor komponenseivel, akkor $1 \leq i, j \leq N$ esetén

$$J_{ij} = \frac{\partial F_x^i(\underset{\sim}{X}^{n+1})}{\partial u_x^{n+1,j}} = \int_\Omega \phi_j \phi_i +$$
$$+ \Delta t \left(\eta_s \left(\nabla_{\underset{\sim}{x}} \phi_j \cdot \nabla_{\underset{\sim}{x}} \phi_i \right) + \left(\phi_j \frac{\partial u_x^{n+1}}{\partial x} + \underset{\sim}{u}_x^{n+1} \cdot \nabla_{\underset{\sim}{x}} \phi_j \right) \phi_i \right) \mathrm{d}\underset{\sim}{x}.$$

J többi elemét hasonlóan számítjuk ki. Ezek után feltéve, hogy a közelítő megoldás k-dik iteráltja a t^n pillanatban $\underset{\sim}{X}_k^n$ és Newton módszerét alkalmazva azt kapjuk, hogy

$$J\underset{\sim}{X}_{k+1}^{n+1} = J\underset{\sim}{X}_k^{n+1} - \underset{\sim}{H}\left(\underset{\sim}{X}_k^{n+1}\right).$$

A megelőző pillanatban kapott megoldást használva kezdeti vektornak – amit $\underset{\sim}{X}_0^{n+1}$ jelöl –, iterációt alkalmazva kiszámítjuk az $\underset{\sim}{X}_k^{n+1}$ vektorokat, amíg csak $\left\| \underset{\sim}{X}_{k+1}^{n+1} - \underset{\sim}{X}_k^{n+1} \right\| < \mathtt{TOL}$ igaz nem lesz. Itt „TOL" egy előre meghatározott tolerancia értéket jelöl.

2.1.2. A Fokker–Planck-egyenlet numerikus megközelítése

Ebben a részben egy végeselem alapú módszert mutatunk be a (7) egyenlet megoldására. Első lépésként, a fentieket követve, felbontjuk az operátort s így két egyenletet kapunk: egy konfigurációs térnek és egy fizikai térnek megfelelőt (ld. a (10), illetve (11) egyenleteket). Vegyük észre, hogy ezek diszkretizált egyenletek, mely a retrográd Euler-módszer eredménye. A Navier–Stokes esethez hasonlóan ezt azért alkalmazzuk, hogy ne legyen korlátozva az időintervallumok hossza. A Fokker–Planck-egyenlet esetében nem előnyös a Crank–Nicolson-módszert használni, mivel a sebességmezőt a $t = t^n$ és nem a $t = t^{n+1/2}$ pillanatban számítjuk ki, s ezért a közelítő megoldások sorozata nem konvergál másodrendben Δt szerint.

Így a háromdimenziós FENE súlyzó modellre koncentrálunk, bár a kétdimenziós FENE esetben hasonló eredmények igazak, valamint a Hooke-féle súlyzó modell esetében is. A háromdimenziós FENE súlyzó modell esetében $\underset{\sim}{F}(\underset{\sim}{q})$-t a (6) egyenlet határozza meg.

A (10) egyenlet gyenge formája alkalmazásával a konfigurációs térbeli feladat a következő: találjuk meg $\tilde{\psi} \in K$-t, amire:

$$\int_D \tilde{\psi} v \, \mathrm{d}\underset{\sim}{q} + \frac{\Delta t}{2\lambda} \int_D \nabla_{\underset{\sim}{q}} \tilde{\psi} \cdot \nabla_{\underset{\sim}{q}} v \, \mathrm{d}\underset{\sim}{q} + \Delta t \int_D \nabla_{\underset{\sim}{q}} \cdot \left(\left(\underset{\approx}{\kappa}^n \underset{\sim}{q} - \frac{1}{2\lambda} \underset{\sim}{F}(\underset{\sim}{q}) \right) \tilde{\psi} \right) v \, \mathrm{d}\underset{\sim}{q} = \int_D \psi^n v \, \mathrm{d}\underset{\sim}{q}, \tag{18}$$

minden $v \in K$ esetén. A (18) egyenletben a diffúziós tag parciális integrálása után megjelenő perem-tag eltűnik a $\psi|_{\partial D} = 0$ Dirichlet-feltétel miatt. Ezen kívül, mivel ebben a modellben feltesszük, hogy a súlyzót alkotó két golyó megkülönböztethetetlen, e parciális differenciálegyenlet megoldása szimmetrikus lesz az origóra. Lásd a [?, ?] cikket e variációs probléma K függvényterének helyes választásáról.

Ahhoz, hogy a (18) egyenletet numerikusan megoldjuk, természetesen adódik, hogy a D tartományt átírjuk gömbi koordináták segítségével, azaz

$$\underset{\sim}{q} = (\rho \cos\theta \sin\varphi, \rho \sin\theta \sin\varphi, \rho \cos\varphi)$$

ahol $(\rho, \theta, \varphi) \in (0, \sqrt{b}) \times (0, 2\pi) \times (0, \pi)$. Ekkor periodikus peremfeltétellel kell dolgoznunk θ szempontjából, mivel a θ változó 2π-vel való elforgatása az identitás operátor. Chauvière és Lozinski munkáját követve (ld. [?|) feltesszük, hogy ψ alakja

$$\psi(\underset{\sim}{x}, \rho, \theta, \varphi, t) = \Psi_0(\rho)\, \alpha(\underset{\sim}{x}, \rho, \theta, \varphi, t), \tag{19}$$

ahol $\Psi_0(\rho) = \left(1 - \rho^2/b\right)^s$ és s egy pozitív állandó. Ha s-t megfelelően választjuk, akkor a behelyettesítés automatikusan homogén Dirichlet-feltételt ad, azaz $\psi = 0$ a ∂D peremen, numerikusan stabil módon, az $\underset{\sim}{F}$ függvény $\rho = \sqrt{b}$ pontban fellépő szingularitása ellenére. [?]-al egyetértésben, empirikusan beláttuk, hogy $s = 2.5$ esetén egy stabil numerikus rendszert kapunk, s ezt az értéket használtuk a ??. részben tárgyalt eredmények esetében.

Most tehát készen állunk arra, hogy implementáljuk a javasolt végeselem módszert egyenletrendszerünk esetében.

Legyen $K_\alpha = \{\alpha : \Psi_0 \alpha \in K\}$ és tegyük fel, hogy $K_{\alpha,h}$ a K_α függvénytér egy véges-dimenziós altere $\{\phi_1, \ldots, \phi_N\}$ bázissal. Az alábbi lépések segítségével a (18) egyenlet térben diszkretizált verzióját kapjuk α függvényében:

(i) végezzük el a (19) helyettesítést;

(ii) legyen $\tilde{\alpha}_h = \sum_j \tilde{\alpha}_j \phi_j$;

(iii) legyen $v = \phi_i$;

(iv) értékeljük ki az integrálokat a $(\rho,\theta,\varphi) \in (0,\sqrt{b}) \times (0,2\pi) \times (0,\pi)$ tartomány felett.

Mindezek után a következő egyenletrendszert kapjuk $i = 1,2,\ldots,N$-re:

$$\sum_{j=1}^{N} \tilde{\alpha}_j \int_0^{\sqrt{b}} \int_0^{2\pi} \int_0^{\pi} \left[\Psi_0 \phi_j \phi_i + \Delta t \left(\underset{\sim}{\nabla}_q \cdot \left(\left(\underset{\approx}{\kappa}^n \underset{\sim}{q} - \frac{1}{2\lambda} \underset{\sim}{F}(q) \right) \Psi_0 \phi_j \right) \phi_i + \right. \right.$$
$$\left. \left. + \frac{1}{2\lambda} \underset{\sim}{\nabla}_q (\Psi_0 \phi_j) \cdot \underset{\sim}{\nabla}_q \phi_i \right) \right] \rho^2 \sin\varphi \, \mathrm{d}\varphi \, \mathrm{d}\theta \, \mathrm{d}\rho =$$
$$= \int_0^{\sqrt{b}} \int_0^{2\pi} \int_0^{\pi} \Psi_0 \alpha^n \phi_i \rho^2 \sin\varphi \, \mathrm{d}\varphi \, \mathrm{d}\theta \, \mathrm{d}\rho. \tag{20}$$

Ezek után a Fokker–Planck-egyenlet fizikai térbeli részét tekintjük (ld. (11)). A (19) egyenletbeli helyettesítés után Ψ_0 kiemelhető az egyenlet minden egyes tagjából, s így a variációs probléma a következőképpen alakul: meg kell találnunk azt az α^{n+1} függvényt K_α-ban, amire

$$\int_\Omega \alpha^{n+1} v \, \mathrm{d}\underset{\sim}{x} + \Delta t \int_\Omega \left(\underset{\sim}{u}^n \cdot \underset{\sim}{\nabla}_x \alpha^{n+1} \right) v \, \mathrm{d}\underset{\sim}{x} = \int_\Omega \tilde{\alpha} v \, \mathrm{d}\underset{\sim}{x}, \tag{21}$$

minden K_α-beli v esetén. Alkalmazva az $\alpha^h = \sum_j \alpha_j \phi_j$ és $v = \phi_i$ végeselem diszkretizációt (ahol $\mathrm{span}\{\phi_1,\ldots,\phi_N\} = K_{\alpha,h} \subset K_\alpha$) azt kapjuk, hogy

$$\sum_{j=1}^{N} \alpha_j^{n+1} \int_\Omega \left(\phi_j + \Delta t \left(\underset{\sim}{u}^n \cdot \underset{\sim}{\nabla}_x \phi_j \right) \right) \phi_i \, \mathrm{d}\underset{\sim}{x} = \int_\Omega \tilde{\alpha} \phi_i \, \mathrm{d}\underset{\sim}{x}, \tag{22}$$

$i = 1,2,\ldots,N$ esetén. Ez egy tisztán hiperbolikus egyenlet Galerkin formája. Eddigi számításaink alapján úgy tűnik (ld. a ??. részt), hogy ez az egyszerű módszer elfogadható. Azonban nagyon előnyös lenne a (21) egyenlet diszkretizálására egy, a klasszikus Galerkin-módszernél stabilabb, módszert használni. Lozinski és Chauvière ecélból spektrál elem alapú áramvonal-diffúziós (Streamline-Upwind Petrov-Galerkin (SUPG)) módszert használtak (ld. [?] és [?] a részletes leírásért).

Vegyük észre, hogy (22) nem függ a konfigurációs térbeli helyzettől. Ez egy nagyon fontos részlet, mivel azt jelenti, hogy a (22) egyenlet szoftveres implementálásakor a rendszer mátrixát időlépésenként csak egyszer kell felállítanunk. Sajnos a (20) egyenletnek nincs meg ez a tulajdonsága, mivel $\underset{\approx}{\kappa}$ függ $\underset{\sim}{x}$-től inhomogén sebességmezők esetén.

Miután ψ-t kiszámítottuk, a Kramers-képletet kell használnunk a polimer $\underset{\approx}{\tau}(\underset{\sim}{x},t)$ extra-feszültségének meghatározására, $(\underset{\sim}{x},t) \in \Omega \times (0,T]$ mellett. Ez egy D feletti integrál kiszámítását jelenti, melyhez szükségünk van a Kramers-képletre α

függvényében:

$$\underset{\approx}{\tau}(\underset{\sim}{x}, t^n) = \zeta_p \left(-\underset{\approx}{I} + \int_0^{\sqrt{b}} \int_0^{2\pi} \int_0^{\pi} \underset{\sim}{w} \otimes \right.$$

$$\left. \otimes \underset{\sim}{w} \frac{\rho^4}{1-\rho^2/b} \Psi_0 \alpha^n(\underset{\sim}{x}, \rho, \theta, \varphi) \sin\varphi \, \mathrm{d}\varphi \, \mathrm{d}\theta \, \mathrm{d}\rho \right), \tag{23}$$

ahol $\zeta_p = \frac{n_p(b+d+2)}{b\lambda} n_p kT$ és $\underset{\sim}{w} = (\cos\theta \sin\varphi, \sin\theta \sin\varphi, \cos\varphi)$.

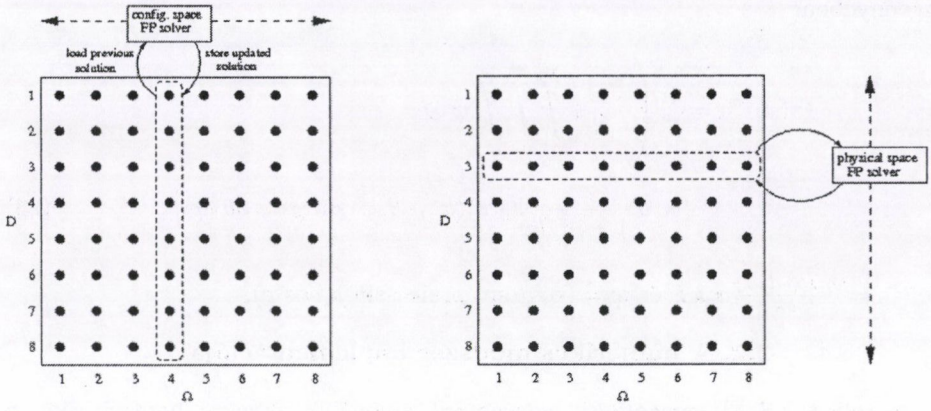
2.2. A numerikus módszer implementációja

A teljes NS–FP megoldási „gépezetet" különálló részekre bonthatjuk: a „Navier–Stokes-megoldóra", a D konfigurációs térbeli „Fokker–Planck-megoldóra", az Ω fizikai térbeli „Fokker–Planck-megoldóra", s végül a Kramer-kifejezésre, amit $\underset{\approx}{\tau}$ kiszámítására használunk. Ezek mindegyikét implementáltuk a `libMesh` nevű végeselem könyvtárt használva. A `libMesh` egy nyitott forráskódú, parallel számítási módszert használó, C++ nyelven íródott szoftver, melyet a „University of Texas at Austin" egyetemen fejlesztettek ki. Számításaink eredményeit a 3. részben összegeztük.

Az implementálás egy lényeges aspektusa, hogy hogyan kezeljük az FP-egyenlet hatváltozós (azaz t-t is figyelembe véve hétváltozós) α megoldását. A kérdést egyszerűen közelítettük meg: az α megoldást minden egyes időbeli lépésben egy olyan mátrixban tároljuk, melynek minden egyes sora egy konfigurációs térbeli rácspontnak megfelelő fizikai térbeli megoldás háromdimenziós keresztmetszetét tartalmazza. Hasonlóan, minden oszlop egy konfigurációs térbeli keresztmetszetet tartalmaz. Ezek után α „frissítése" az egyes keresztmetszetek egymás utáni frissítésével történik.

Az alábbi lista a teljes számítási eljárás egy pontosabb leírását adja.

1. Kezdetben a rendszert egyensúlyi helyzetbe állítjuk az $\underset{\sim}{u}(\underset{\sim}{x}, 0) = \underset{\sim}{0}$ és $\psi(\underset{\sim}{x}, \underset{\sim}{q}, 0) = \psi_{eq}(\underset{\sim}{q})$ választással. Itt $\psi_{eq}(\underset{\sim}{q}) = C\left(1 - |\underset{\sim}{q}|^2/b\right)^{b/2}$ és C egy normalizációs állandó (ld. [5]). Legyen még $\underset{\approx}{\tau}(\underset{\sim}{x}, 0) = \underset{\approx}{0}$, mivel egyensúlyi helyzetben a polimer extra-feszültségi tenzora eltűnik.
2. Vegyünk egy nem nulla beáramlási peremfeltételt az $\underset{\sim}{u}$ további peremfeltételeinek megfelelő beállításával, és frissítsük a sebességmezőt a 2.1.1. bekezdésben tárgyalt Navier–Stokes-megoldó segítségével.
3. Frissítsük α-t D szempontjából úgy, hogy a fizikai tér hálója rácspontjai felett iterálunk, és a (20) egyenlet segítségével frissítjük a konfigurációs térbeli keresztmetszeteket.
4. Frissítsük α-t Ω szempontjából minden D-beli rácspontra, a (22) egyenletben megadott Galerkin-implementáció segítségével. Mint ahogy korábban



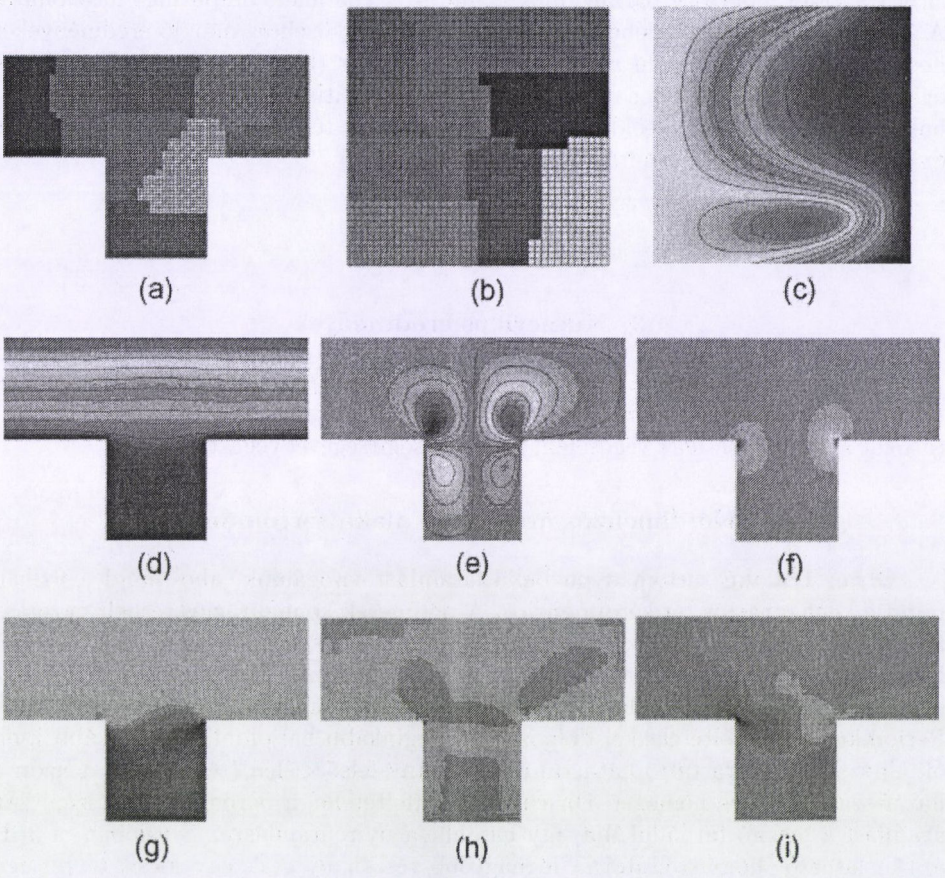
2. ábra. Ez az ábra a Fokker–Planck-egyenlet megoldásának operátor felbontásos módszerét illusztrálja. Minden egyes időbeli lépésben először (a) frissítünk minden egyes konfigurációs térbeli keresztmetszetet, aztán (b) frissítünk minden egyes fizikai térbeli keresztmetszetet.

is említettük, ebben az esetben elég a rendszer mátrixát egyszer felállítani minden egyes időintervallumra. Ez azt eredményezi, hogy a fizikai térbeli frissítések lényegesen kevesebb CPU időt igényelnek, s a különbség nő a feladat méretének növekedésével.

5. Frissítsük $\underset{\approx}{\tau}$-t a frissített Fokker–Planck-megoldás alapján, a (23) egyenlet integrálja kiszámításával, melyhez Gauss-kvadratúrát használunk.

6. Frissítsük $\underset{\sim}{u}$-t – erre a frissített $\underset{\approx}{\tau}$ polimer extra-feszültségi tenzor van hatással. Térjünk vissza a 3. lépéshez, és ismételjük a 3–6 folyamatot, amíg valamely leállási feltétel, mint például $\frac{\|\underset{\sim}{u}^{n+1}-\underset{\sim}{u}^{n}\|_\infty}{\Delta t} < \texttt{TOL}$, igaz nem lesz.

A fenti algoritmus legnagyobb része az α leképezés konfigurációs térbeli frissítésével telik. Nagyon fontos volna e lépés optimalizálása. Lozinski és Chauvière [23]-ban egy olyan „gyors Fokker–Planck-megoldót" dolgoztak ki, melynek köszönhetően algoritmusuk teljes CPU időigénye több mint 60-as szorzóval javult. Ezt a (18) egyenletben levő, spektrális módszerből adódó, kifejezés átrendezésével érték el, mellyel megmutatták, hogy számos mátrix és inverzük előre kiszámítható, s újra és újra felhasználható minden egyes megoldás során, ezzel drasztikusan csökkentve a konfigurációs térbeli megoldásokhoz szükséges munka mennyiségét. Ez azonban nem alkalmazható végeselem megközelítés során; viszont a végeselem módszer egyik nagy előnye, hogy segítségével szinte kizárólag ritka mátrix algebrát tudunk



3. ábra. Az (a) és (b) ábrán a 8 processzor közötti terhelés megoszlása látható, a fizikai és konfigurációs térbeli hálóknak megfelelően. A csatorna Ω-beli fő részének egy x pontjához tartozó konfigurációs térbeli α keresztmetszete a (c) ábrán látható. Ez a keresztmetszet hasonlít a nyíróáramláshoz tartozó FP-egyenlet megoldására, mint ahogy az várható volt. A (d), (e) és (f) ábrákon u_x, u_y illetve p függvények stacionárius megoldásait láthatjuk, $\underset{\approx}{\tau}$ komponenseit pedig az utolsó sorban (azaz τ_{xx}-t a (g), τ_{xy}-t a (h) és τ_{yy}-t az (i) ábrán).

alkalmazni. Mivel ritka mátrix inverze általában teljesen kitöltött mátrix, nem kívánatos ezeket explicite kiszámítani.

Így tehát más megoldást kell találnunk a Fokker–Planck-megoldó felgyorsítására. Egy lehetőség párhuzamos számítások (parallel computing) használata. A `libMesh` könyvtár ezt lehetővé teszi, és ez a megközelítés már jó eredményeket hozott nagy nagyságrendű számítások elvégzésekor (ld. a 3. részt). Egy másik ötlet (melyet a 4. részben tárgyalunk) az, hogy ritkított térhálót (sparse grid) használunk azért, hogy csökkentsük a konfigurációs térbeli rács szabadságfokát, s ezáltal csökkentsük a számításokhoz szükséges időt.

3. Numerikus eredmények

Ebben a részben bemutatjuk a polimer áramlások numerikus modelljei alapján kapott eredményeinket mind két-, mind pedig háromdimenziós esetben. A számításokat a 2. részben leírt végeselem módszer segítségével végeztük.

3.1. Kétdimenzós áramlás T alakú tartományban

Olyan T alakú tartományon belüli áramlást vizsgálunk, ahol mind a fizikai, mind a konfigurációs tér kétdimenziós. A polimerek konfigurációs térbeli mozgása általában nem korlátozódik ugyanarra a kétdimenziós síkra, még a fizikai térbeli lamináris sebességmező esetén sem, úgyhogy kétdimenziós konfigurációs térrel dolgozni nem reális (ld. [9]). Most mégis ezt az esetet vizsgáljuk egyszerűsítés céljából. Periodikus peremfeltételekkel dolgozunk Ω leginkább bal oldali és leginkább jobb oldali szélén, jobbra tartó határral a tartomány felső szélén (tehát $u_x = 1$ azon a határ-szakaszon) és homogén Dirichlet-peremfeltétellel a perem többi részén. Az áramlást a mozgó fal indukálja, így hasonlít a nyíróáramlásra. S valóban, a 3(d) ábrán látható, hogy a csatorna legnagyobb részén u_x csak egy kicsit tér el egy egyszerű nyíróáramlás sebességmezejétől. Itt a $\lambda = 1, b = 10, \eta_p = 1.439$ és $\eta_s = 1$ paramétereket használtuk.

Ezt a szimulációt a Texas Advanced Computing Centre számítástechnikai központ (`http://www.tacc.utexas.edu`) Lonestar nevű parallel számítógépe 8 processzán futtattuk. 100 időintervallumot kellett használnunk, $\Delta t = 0.05$ léptékkel ahhoz, hogy elérjük a 3. ábrán látható stacionáris megoldást. A számítás 81 másodpercet vett igénybe időintervallumonként, melynek 71%-a konfigurációs térbeli, míg 23%-a fizikai térbeli frissítésekkel telt.

Folytonos bázisfüggvényekkel dolgoztunk. A fizikai térbeli háló 1024 bi-kvadratikus elemből (4257 rácspontból) állt, míg a konfigurációs téré 400 bi-kvadrikus elemből (1681 rácspontból).

3.2. Henger körüli kétdimenziós áramlás

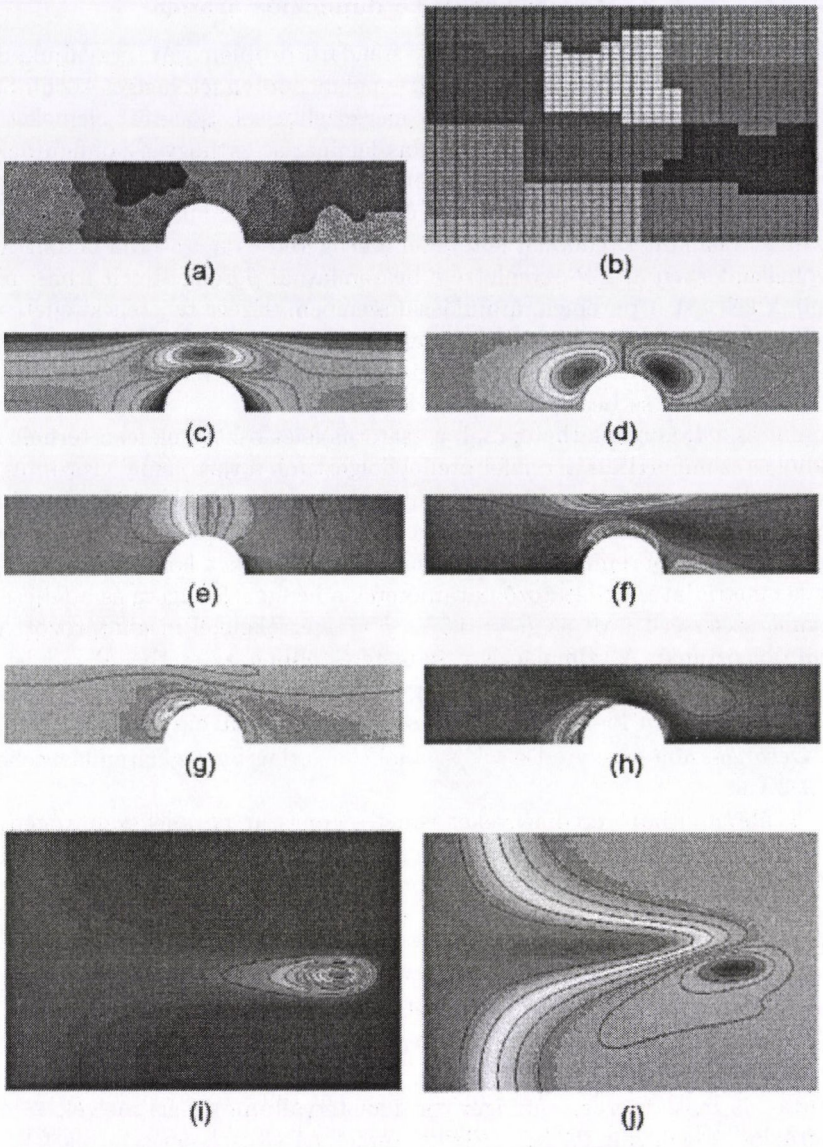
A következőkben a szakirodalom egy standard problémáját tárgyaljuk, melyet gyakran használnak viszonyítási alapnak – polimer-folyadék henger körüli Stokes-áramlását; (ld. [9, 23] determinisztikus megközelítéssel, spektrál elemeket használva). Ismét feltesszük, hogy az áramlás lamináris, és hogy a konfigurációs tér kétdimenziós. A sebesség szempontjából a következő megszorításokkal éltünk: a bal és jobb peremen u_x esetében parabolikus ki- és beáramlási peremfeltételekkel dolgoztunk, valamint tapadó fal feltétellel a hengeren és a csatorna falain. Ahhoz, hogy a fizikai térben az FP-egyenletre a beáramlásnál peremfeltételt írhassunk elő, ismernünk kell ψ-t a peremen, áramlással szemben. Ezért teljesen kifejlett áramlással dolgozunk a perem közelében, azaz a beáramlási peremfeltétel parabolikus formájával. Így a vonatkozó valószínűségi sűrűségfüggvény az adott peremfeltételek mellet kiszámítható és beáramoltatható Ω-ba.

Amint az a 4. ábrán látható, csak a csatorna felét osztottuk fel a térháló szempontjából és szimmetrikus peremfeltétellel dolgoztunk a tartomány vízszintes szimmetriatengelye mentén. Ezt azért tehettük meg, mert henger körüli Stokes-áramlás mindenképpen szimmetrikus. A félhenger keresztmetszetét képező félkör mentén homogén Dirichlet-peremfeltételt alkalmaztunk $\underset{\sim}{u}$ mindkét komponensére. A fizikai tér geometriáját meghatározó paraméterek a henger R sugara és a fél-csatorna y-irányú h szélessége. Mi az $R = 0.5$ és $h = 1$ értékekkel meghatározott tartományon dolgoztunk. A szimulációs paraméterek ebben az esetben $\lambda = 1$, $b = 20$, $\eta_p = 1.439$ és $\eta_s = 1$ voltak. Ahhoz, hogy eredményeinket össze lehessen hasonlítani a [23]-beliekkel a ki- és beáramlási sebességprofilokat úgy választottuk, hogy az ún. *Deborah-szám* (melyre $\mathrm{De} = \frac{\lambda \bar{U}}{R}$, ahol $\bar{U}$ az átlag be- és kiáramlási sebesség) éppen 1.2 volt.

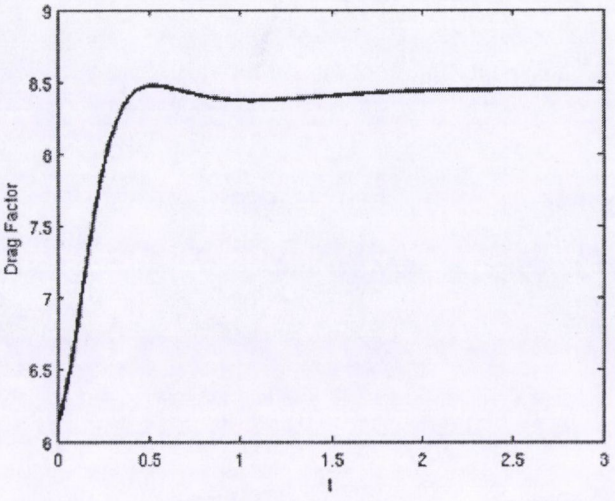
A 4. ábrán látható eredményeket ismét a Lonestar parallel számítógép segítségével kaptuk, ebben az esetben 10 processzort használva. A fizikai és konfigurációs térbeli terhelés megoszlása a 4-es ábrán látható. 300 időintervallumot kellett használnunk, $\Delta t = 0.01$ léptékkel ahhoz, hogy elérjük a stacionáris megoldást. Triangulált fizikai térbeli hálóval dolgoztunk a `Triangle` nevű háló-generáló szoftver segítségével (ld. [30]), mely 1062 elemből (2239 rácspontból) állt. Folytonos, darabonként kvadratikus bázisfüggvényeket alkalmaztunk a sebességvektor komponensei esetében és folytonos, darabonként lineárisakat a nyomásra. A fizikai és konfigurációs tér halói 400 bi-kvadratikus elemből (1681 rácspontból) álltak. Ez a számítás 49 másodpercet vett igénybe időintervallumonként, melyek 71%-át a konfigurációs térbeli, míg 24%-át a fizikai térbeli frissítések emésztették fel.

E viszonyítási alapként használt áramlás numerikus modelljei közötti különbségek kimutatására gyakran használják a henger körüli ún. *közegellenállási tényező (drag factor)* makroszkopikus mennyiséget. Ezt F^* jelöli; pontos definíciója megtalálható [23]-ben.

A fenti eredményekhez tartozó F^* értékeket kiszámítottuk $t \in [0, 3]$ esetében és az így kapott adatokat az 5. ábrán foglaltuk össze. Ez megegyezik a [23] cikkben közölt eredményekkel $\mathrm{De} = 1.2$ esetén.



4. ábra. Az (a) és (b) ábrán a 10 processzor közötti terhelés megoszlása látható. A stacionáris megoldáshoz tartozó sebesség u_x és u_y komponensei, valamint a nyomás a (c), (d), illetve (e) ábrán, a τ_{xx}, τ_{xy} és τ_{yy} komponensek az (f), (g), illetve (h) ábrán láthatóak. Annak illusztrálására, hogy az FP-egyenlet konfigurációs térbeli megoldása hogyan függ a fizikai térbeli helyzettől, az (i) és (j) ábrán α két különböző D-beli keresztmetszete látható.



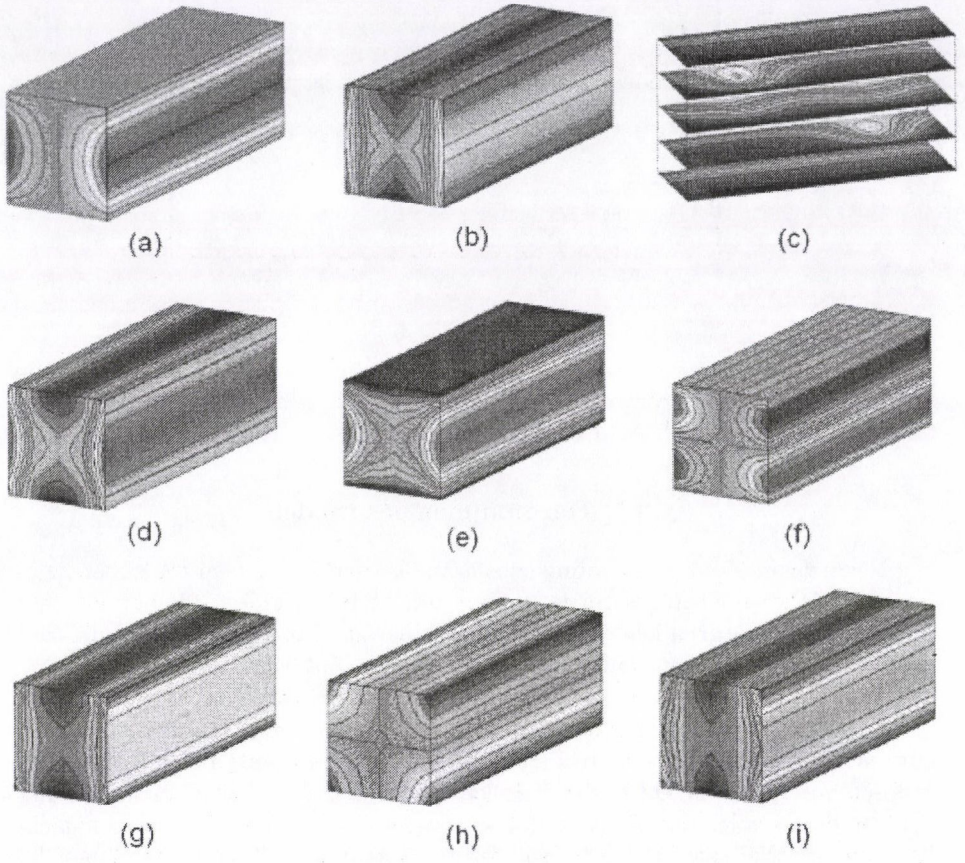
5. ábra. Az közegellenállási tényező (drag factor) idő szerinti függvénye a henger körüli áramlás esetében.

3.3. Háromdimenziós modell

Végül tekintsük a háromdimenziós áramlás esetét. A fizikai tér ebben az esetben egy téglatest alakú csatorna, s az áramlást ismét mozgó fal peremfeltételek indukálják. A csatorna két végén periodikus peremfeltételekkel dolgozunk. Ebben az esetben csak egy nem nulla sebességkomponensünk van és ez u_y. A $\lambda, b, \eta_s, \eta_p$ szimulációs paraméterek ugyanazok, mint a 3.1. részben.

A konfigurációs tér hálója $10 \times 10 \times 10$ darab téglatestből állt; folytonos, darabonként tri-kvadratikus bázisfüggvényeket alkalmazunk a sebesség esetében, és folytonos, darabonként tri-lineárisakat a nyomásra. A fizikai tér hálója meglehetősen durva volt: mindössze $4 \times 4 \times 4$ darab téglatestből állt. A szimulációt a Lonestar parallel számítógép 8 processzorán futtattuk, 60 időintervallumot használva $\Delta t = 0.05$ léptékkel, s a számítás átlagosan 129.9 másodpercet vett igénybe időintervallumonként. A számítási idő megoszlása hasonlít a kétdimenziós esethez: a CPU idő 74%-a konfigurációs térbeli, míg 20%-a fizikai térbeli frissítésekkel telt. Az α mezőt és a feszültségi tenzor komponenseit a 6. ábrán mutatjuk be.

A teljes háromdimenziós (azaz háromdimenziós fizikai és háromdimenziós konfigurációs térrel) való számítás sok kihívással jár. Ismereteink szerint ez a jelen munka az első olyan, ahol determinisztikus módszert használnak teljes háromdimenziós modell kidolgozására (Lozinski és Chauvière háromdimenziós konfigurációs és kétdimenziós fizikai teret vettek [8]-ban). Szándékunkban áll ennél összetettebb háromdimenziós áramlások modellezése is; ez azonban nem egyszerű fel-



6. ábra. Az ábrán a következő stacionáris megoldások láthatók: (a) u_y (u_x és u_z azonosan nulla, ezért ezeket nem ábrázoljuk), (b) p, (c) α keresztmetszete a konfigurációs térben, (d) τ_{xx}, (e) τ_{xy}, (f) τ_{xz}, (g) τ_{yy}, (h) τ_{yz}, (i) τ_{zz}.

adat. Az elsődleges nehézség abban áll, hogy a térháló szabadságfokának növekedésével egyre több egymás utáni frissítésre van szükség mind a konfigurációs, mind a fizikai térben.

A számításokhoz használt processzorok számának növelése legfeljebb állandó értéken tartja az egyes frissítésekhez szükséges időt (ahogy a térháló finomodik), de mivel több frissítésre van szükség – sőt, lényegesen többre, mivel három dimenzióban a szabadságfok gyorsan nő –, a számításhoz szükséges teljes CPU idő is nőni fog.

A későbbiekben szándékunkban áll különféle numerikus módszereket megvizsgálni abból a szempontból, hogy ezen a problémán javítsunk.

4. További tervek, végszó

Munkánkat számos területen fogjuk folytatni azért, hogy továbbfejlesszük az előzőekben bemutatott ötleteket. Ezeket tárgyaljuk ebben a részben.

Mint azt a 2. részben megjegyeztük, az általunk bemutatott determinisztikus mikro-makro algoritmushoz szükséges számítási idő legnagyobb részét az FP-egyenlet konfigurációs térbeli részének megoldása viszi el. Ezért megközelítésünk hatásfokának növelése érdekében szeretnénk az eljárás ezen részének optimalizálására koncentrálni. A [32] és [29] munkákat követve, erre egy lehetőség a stabilizált ritkított végeselem módszerek használata lenne. Ritkított végeselem módszerek esetében az approximációelméleti eredmények (ld. [7] és [29]) azt mutatják, hogy e módszerek lehetőséget nyújtanak egy bizonyos pontossági szint elérésére kisebb szabadságfok mellett, mint teljes szorzatháló esetén, bár ekkor szigorúbb simasági feltételekre van szükség. További előny, hogy a szabadságfok csökkenése nő a térháló dimenziójának növekedésével. Amennyiben ritkított térhálók alkalmazása lehetővé teszi a konfigurációs térbeli szabadságfok csökkentését, akkor ez jelentős megtakarítást jelent, mivel nemcsak, hogy csökkentjük minden egyes konfigurációs térbeli frissítésnél a feladat méretét, de csökkentjük a szükséges fizikai térbeli megoldás keresések számát is.

Egy másik fontos kérdés, hogy kersztülvihető-e determinisztikus megközelítést használnunk olyan modellek esetében, ahol a konfigurációs tér dimenziója több mint három – mint például a Rouse–Zimm-lánc polimer modellen alapulóknál. Az a véleményünk, hogy ritkított térhálók alkalmazása ezt lehetővé teszi. Valóban, [11]-ben számításokat végeztek, éppen ezt a megközelítést használva, olyan egyszerű homogén nyíróáramlás esetén, ahol a konfigurációs tér dimenziója a hatot is elérte. A leközölt eredmények alapján úgy tűnik, hogy ritkított térhálók alkalmazása jelentősen növeli a hatásfokot a teljes tenzorszorzat háló használatához képest. Ritkított térhálók alkalmazása golyó-rugó típusú polimerlánc modellek esetén egy olyan érdekes téma, amit meg szeretnénk vizsgálni további munkánk során.

Kutatásaink egy másik fontos iránya algoritmusunk matematikai elemzésének finomítása. Egy dolog, amit tisztázni kell, az az operátor felbontás utáni FP-egyenlet gyenge formájának felírásakor használt függvényterek megválasztása. A 2. részben a függvénytereket K-val és M-mel jelöltük és semmit sem mondtunk ezek struktúrájáról. Ezen kívül szeretnénk a determinisztikus algoritmus konvergenciájával is foglalkozni. Lozinski és munkatársai (például [8,9,23]-ben) az általuk tárgyalt determinisztikus megközelítés esetében megfigyelték, hogy a megoldás konvergál a térháló finomításával. Szeretnénk ezt a kérdést nagyobb matematikai szigorral megvizsgálni. Ilyen irányú kutatásainkat a [19] munkánk összegezi.

Szándékunkban áll még kielemezni, milyen hatással bír $\psi(\underset{\sim}{x}, \cdot, t)$ pontossága a releváns makroszkopikus mennyiségekre. Mivel a ψ függvényt a $\underset{\sim}{q}$ változó szerint átlagoljuk ahhoz, hogy kiszámítsuk $\underset{\sim}{\tau}$-t, valószínű, hogy ψ pontossága $\underset{\sim}{q}$-ra nézve kevésbé fontos, mint $\underset{\sim}{x}$-re és t-re nézve. Amennyiben sikerül kidolgoznunk a mikroszkopikus és makroszkopikus változók diszkretizációjainak (esetleg adaptív) finomításai közötti pontos összefüggést, akkor lehetővé válna a számításhoz szükséges eszközök és erőforrások (pl. idő, tárhely) hatékonyabb beosztása.

Jelen cikkünkben megvizsgáltunk egy, oldott polimer-folyadékok dinamikáját leíró, többskálájú Navier–Stokes Fokker–Planck-modell numerikus megoldására javasolt végeselem alapú módszert. Bemutattunk néhány számítási eredményt azért, hogy módszerünk hatékonyságát demonstráljuk. A fentebb röviden tárgyalt további kutatási tervek megvalósításakor szándékunkban áll a determinisztikus megközelítés maximális kihasználása azért, hogy lássuk, felveszi-e a versenyt az eddig kidolgozott sztochasztikus módszerekkel a különféle polimer-folyadékok modellezésében.

Hivatkozások

[1] Barrett, J.W. and Schwab, Ch. and Süli, E.: *Existence of global weak solutions for some polymeric flow models*, Math. Models and Methods in Applied Sciences **15**, 3 (2005), 939–983.

[2] Barrett, J.W. and Süli, E.: *Existence of global weak solutions to some regularized kinetic models of dilute polymers*, SIAM J. Multiscale Modeling and Simulation **6**, 2 (2007), 506–546, To appear.

[3] Biller, P. and Petruccione, F.: *The flow of dilute polymer solutions in confined geometries: a consistent numerical approach*, J. Non-Newtonian Fluid Mech. **25**, (1987), 347–364.

[4] Bird, R.B. and Curtiss, C.F. and Armstrong, R.C and Hassager, O.: *Dynamics of Polymeric Liquids, Volume* **1**, *Fluid Mechanics*, second ed., John Wiley and Sons, (1987).

[5] Bird, R.B. and Curtiss, C.F. and Armstrong, R.C and Hassager, O.: *Dynamics of Polymeric Liquids, Volume* **2**, *Kinetic Theory*, second ed., John Wiley and Sons, (1987).

[6] Brenner, S.C. and Scott, L.R.: *The Mathematical Theory of Finite Element Methods*, second ed., Springer (2002).

[7] BUNGARTZ, H.J. AND GRIEBEL, M.: *Sparse Grids*, Acta Numerica (2004), 1–123.

[8] CHAUVIÈRE, C. AND LOZINSKI, A.: *Simulation of complex viscoelastic flows using Fokker-Planck equation: 3D FENE model*, J. Non-Newtonian Fluid Mech. **122**, (2004), 201–214.

[9] CHAUVIÈRE, C. AND LOZINSKI, A.: *Simulation of dilute polymer solutions using a Fokker-Planck equation*, Computers and Fluids **33**, (2004), 687–696.

[10] CHAUVIÈRE, C. AND OWENS, R.G.: *A new spectral element method for the reliable computation of viscoelastic flow*, Comput. Methods Appl. Mech. Eng. **190**, (2001), 3999–4018.

[11] DELAUNAY, P. AND LOZINSKI, A. AND OWENS, R.G.: *Sparse tensor-product Fokker-Planck-based methods for nonlinear bead-spring chain models of dilute polymer solutions*, CRM Proceedings and Lecture Notes, (2006), Preprint.

[12] FAN, X.J.: *Viscosity, first normal stress coefficient and molecular stretching in dilute polymer solutions*, J. Non-Newtonian Fluid Mech. **17**, (1985), 125–144.

[13] FAN, X.J.: *Molecular models and flow calculations: II. Simulation of steady planar flow*, Acta Mechanica Sinica **5**, (1989), 216–226.

[14] GROSSO, M. AND MAFFETTONE, P.L. AND HALIN, P. AND KEUNINGS, R. AND LEGAT, V.: *Flow of nematic polymers in eccentric cylinder geometry: influence of closure approximations*, J. Non-Newtonian Fluid Mech. **94**, (2000), 119-134.

[15] HALIN, P. AND LIELENS, G. AND KEUNINGS, R. AND LEGAT, V.: *The Lagranian particle method for macroscopic and micro-macro viscoelastic flow computations*, J. Non-Newtonian Fluid Mech. **79**, (1998), 387–403.

[16] HULSEN, M.A. AND VAN HEEL, A.P.G. AND VAN DEN BRULE, B.H.A.A.: *Simulation of viscoelastic flows using Brownian configuration fields*, J. Non-Newtonian Fluid Mech. **70**, (1997), 79–101.

[17] KEUNINGS, R.: *A survey of computational rheology*, XIIIth International Congress on Rheology, (Cambridge, UK, August 2000), Available at `http://www.mate.tue.nl/~hulsen`.

[18] KEUNINGS, R.: *Micro–Macro Methods for the Multiscale Simulation of Viscoelastic Flow Using Molecular Models of Kinetic Theory*, Rheology Review (2004), 67–98.

[19] KNEZEVIC, D. AND SÜLI, E.: *Spectral Galerkin approximation of Fokker–Planck equations with unbounded drift*, Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford OX1 3QD, Numerical Analysis Technical Report Series, 2007/16, (2007).

[20] KRAMERS, H.A.: *The viscosity of macromolecules in a streaming fluid*, Physica **11**, 1 (1944), .

[21] LASO, M. AND ÖTTINGER, H.C.: *Calculation of viscoelatic flow using molecular models: the CONNFFESSIT approach*, J. Non-Newtonian Fluid Mech. **47**, (1993), 1–20.

[22] LOZINSKI, A.: *Spectral methods for kinetic theory models of viscoelastic fluids*, École Polytechnique Fédérale de Lausanne, (2003).

[23] LOZINSKI, A. AND CHAUVIÈRE: *A fast solver for Fokker-Planck equation applied to viscoelastic flows calculation: 2D FENE model*, Journal of Computational Physics **189**, (2003), 607–625.

[24] LOZINSKI, A. AND CHAUVIÈRE, C. AND FANG, J. AND OWENS, R.G.: *Fokker-Planck simulations of fast flows of melts and concentrated polymer solutions in complex geometries*, J. Rheology **47**, (2003), 535–561.

[25] NAYAK, R.: *Molecular simulation of liquid crystal polymer flow: a wavelet-finite element analysis*, MIT (1998).

[26] OLDROYD, J.G.: *On the formulation of rheological equations of state*, Proc. Roy. Soc. London **A200**, (1950), 523-541.

[27] ÖTTINGER, H.C.: *Stochastic processes in polymeric fluids*, Springer, (1996).

[28] ROUSE, P.E.: *A theory of the linear viscoelastic properties of dilute soluns of coiling polymers*, J. Chem. Phys. **21**, (1953), 1272–1280.

[29] SCHWAB, C. AND SÜLI, E. AND TODOR, R.-A.: *Sparse finite element approximation of high-dimensional transport-dominated diffusion problems*, Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford OX1 3QD, Numerical Analysis Technical Report Series 2007/4, (2007), .

[30] SHEWCHUK, J. R.: *Delaunay refinement algorithms for triangular mesh generation*, Computational Geometry: Theory and Applications **22**, (2002), 21–74.

[31] STEWART, W.E. AND SØRENSEN, J.P.: *Hydrodynamic interaction effects in rigid dumbbell suspensions: II. Computations for steady shear flow*, Trans. Soc. Rheol. **16**, (1972), 1–13.

[32] SÜLI, E.: *Finite element approximation of high-dimensional transport-dominated diffusion problems;* Oxford University Computing Laboratory, NA-05/19, (2005).

[33] WARNER, H.R.: *Kinetic Theory and Rheology of Dilute Suspensions of Finitely Extendible Dumbbells*, Ind. Eng. Chem. Fundamentals, (1972), 379–387.

[34] ZIMM, B.H.: *Dynamics of polymer molecules in dilute solution: viscoelasticity, flow birefringence and dielectric loss*, J. Chem. Phys. **24**, (1956), 269–278.

DAVID KNEZEVIC
Oxford University Computing Laboratory
Wolfson Building, Parks Road
Oxford OX1 3QD, U.K.

SÜLI ENDRE
Oxford University Computing Laboratory
Wolfson Building, Parks Road
Oxford OX1 3QD, U.K.
Endre.Suli@comlab.ox.ac.uk

FINITE ELEMENT METHODS FOR DETERMINISTIC SIMULATION OF POLYMERIC FLUIDS

DAVID KNEZEVIC, EDRE SÜLI

In this work we consider a finite element approach to solving the coupled Navier-Stokes (NS) and Fokker-Planck (FP) multiscale model that describes the dynamics of dilute polymeric fluids.

Deterministic approaches such as ours have not received much attention in the literature because they present a formidable computational challenge due to the fact that the analytical solution of the Fokker-Planck equation may be a function of a large number of independent variables. For instance, to simulate a 3-dimensional flow using the dumbbell model for polymers one must solve the coupled NS-FP system in which the Fokker-Planck equation is posed in a 6-dimensional domain. First we discuss the physical and mathematical foundations of the NS-FP model, then we develop our deterministic finite element based approach in detail. Numerical results, obtained using parallel computation, are presented to demonstrate the efficacy of our approach.

Alkalmazott Matematikai Lapok 26 (2009), 179-191.

A STOKES-FELADAT ÉS A CROUZEIX–VELTE-FELBONTÁS

STOYAN GISBERT

Ebben a cikkben szeretnénk áttekintést adni saját és tanítványaink eredményei felett: a Stokes-rendszerre vonatkozó elsőfajú peremérték feladat esetén olyan diszkretizációkat kutatunk, amelyekre létezik egy 3 altérből álló speciális ortogonális felbontás. Ezzel nemcsak az analitikus megoldásnak egy jellegzetes tulajdonságát őrizzük meg, hanem lényegesen javul a diszkretizált egyenletek iteratív megoldási módszereinek hatékonysága. Más feladatokra is átvihető ez a technika.

1. A fizikai feladat

Az összenyomhatatlan közegek áramlását a következő matematikai modell írja le:

$$\frac{\partial \vec{u}}{\partial t} + \mathbf{D}(\vec{u})\vec{u} + \operatorname{grad} p = \nu \Delta \vec{u} + \vec{f}, \tag{1}$$

$$\operatorname{div} \vec{u} = 0, \qquad x \in \Omega \in \mathbb{R}^d, \tag{2}$$

$$\vec{u} = 0, \qquad x \in \Gamma, \quad 0 < t \leq T, \tag{3}$$

ahol t az idő, ν a (kinematikus) viszkozitás, $\vec{u}$ a közeg sebesség vektora, a $\mathbf{D}(\vec{u})\vec{u}$ „inerciális tag" olyan vektor, amelynek i-edik komponense $\operatorname{div}(u_i\vec{u})$, továbbá p a (kinematikus) nyomás, $\rho\vec{f}$ a külső erők sűrűségét leíró vektor, és itt ρ a sűrűség. A feladat dimenziója $d = 2, 3$, Ω az áramlási tartomány, amelynek Γ peremétől feltesszük, hogy poligonális és Lipschitz-folytonos.

A fenti (1)–(2) rendszert (összenyomhatatlan) Navier–Stokes-egyenletrendszernek hívjuk, ebből (1) az impulzus megmaradását adja (a konstansnak vett ρ-val való osztása után), és (2) a tömegmegmaradási egyenlet, ugyancsak azután, hogy $\rho = \text{const}$-nak megfelelően egyszerűsítettük.

A (3) peremfeltétel nem az egyetlen lehetséges, ld. Gáspár Csaba cikkét ebben a kötetben. Viszont a 8. pont kivételével minden alábbi eredmény erre az esetre vonatkozik. Fizikailag arról van szó, hogy az Ω áramlási tartomány fala zárt és nem mozog maga, és ehhez a falhoz tapad le a folyadék.

(3)-on kívül még további mellékfeltételek szükségesek: mint kezdetiértéket az $\vec{u}(0,x)$-et egész Ω-ban kell előírni, ezenkívül a nyomás csak egy konstans erejéig van meghatározva: ha p megoldás, akkor $p+c$ is az, tetszőleges c konstanssal.

Ha (1)–(3)-ban csak az időt diszkretizáljuk, azaz szemidiszkretizációt hajtunk végre, akkor jutunk a következő alakra:

$$\begin{aligned} \frac{\vec{u}^j}{\tau} - \nu\Delta\vec{u}^j + \operatorname{grad} p^j &= \frac{\vec{u}^{j-1}}{\tau} - \mathbf{D}(\vec{u}^{j-1})\vec{u}^{j-1} + \vec{f}^j, \\ \operatorname{div}\vec{u}^j &= 0, \qquad x \in \Omega, \\ \vec{u}^j &= 0, \qquad x \in \Gamma,\ 0 < j \le m,\ m = T/\tau, \end{aligned}$$

ahol $\vec{u}^j$ a $t_j := j\tau$ időponthoz tartozó sebesség approximációja. Mint kezdetiértéket az $\vec{u}^0 = \vec{u}(x,0)$-t kell előírni minden $x \in \Omega$-ra. Amit ezután minden j-re meg kell oldani, az már egy Stokes-típusú feladat $\vec{u}^j$ és p^j meghatározására.

Ehhez közel áll a klasszikus Stokes-feladat:

$$-\Delta\vec{u} + \operatorname{grad} p = \vec{f}, \tag{4}$$

$$\operatorname{div}\vec{u} = g, \qquad x \in \Omega, \tag{5}$$

$$\vec{u} = 0, \qquad x \in \Gamma. \tag{6}$$

Itt a ν viszkozitást 1-nek vettük; tipikusan $g = 0$. (1)-hez képest eltűnt a $\frac{\partial\vec{u}}{\partial t}$ (mert most az áramlás nem változik az idővel) és az inerciális tag (mert ez kicsi sebességek és kicsi deriváltjaik mellett másodrendű).

A (4)–(6) rendszer tehát olyan stacionáriusan áramló közeg leírását adja, amelynek sebessége kicsi. Ilyen helyzet mégis gyakorlatilag is érdekes lehet: pl. egy ipari baleset révén a levegőbe jutott szennyeződes leginkább akkor súlyos probléma, ha a szélsebességek kicsik.

(4)–(6)-ból a nyomás egyértelműen meg van határozva, megint csak egy tetszőleges additív konstanst nem számítva. Úgy, mint (1)–(3) esetén, ez arra mutat rá, hogy a feladat szinguláris. A megoldhatósági feltétel megkapható (5) integrálásával, felhasználva (6)-ot :

$$\int_\Omega g(x)\mathrm{d}x = \int_\Omega \operatorname{div}\vec{u}\mathrm{d}x = \int_\Gamma \vec{u}\cdot\vec{n}\mathrm{d}s = 0.$$

Ez azt jelenti, hogy g-t nem lehet tetszőlegesen megadni, hanem az Ω feletti integrálja nulla kell, hogy legyen. Fizikailag ez a feltétel azt fejezi ki, hogy a zárt falban áramló közeg esetén csak annyi tömeg keletkezhet, amennyi eltűnik: az összmérlegnek nullának kell lennie.

2. A gyenge megoldás

A (4)–(6) feladat úgynevezett gyenge megoldásával foglalkozunk a továbbiakban, amikor is keresett $\vec{u} \in V$, $p \in Q$ a szokásos V, Q Szoboljev-terekkel [8]:

$$V := \left(H_0^1(\Omega)\right)^d, \qquad Q := L_2(\Omega),$$

és legyen

$$L_{2,0}(\Omega) := \left\{ q \in L_2(\Omega), \int_\Omega q(x)\mathrm{d}x = 0 \right\}.$$

Ezek Hilbert-terek, amelyekhez az alábbi ismert skalárszorzatok tartoznak:

$$(\vec{u},\vec{v})_1 := \int_\Omega \sum_{i=1}^{d}\sum_{j=1}^{d} \frac{\partial u_i}{\partial x_j}\frac{\partial v_i}{\partial x_j}\mathrm{d}x, \quad \vec{u},\vec{v} \in V,$$

$$(p,q)_0 := \int_\Omega p(x)q(x)\mathrm{d}x, \quad p,q \in Q,$$

és a következő normák:

$$\|\vec{u}\|_1 := (\vec{u},\vec{u})_1^{1/2}, \quad \|p\|_0 := (p,p)_0^{1/2}.$$

Ezen jelölések segítségével felírható a következő variációs feladat, amely a gyenge megoldást definiálja (és egyben ez a végeselem módszer alapja is):

Keresett tehát $\vec{u} \in V$, $p \in Q$ úgy, hogy

$$a(\vec{u},\vec{v}) + b(\vec{v},p) := (\vec{f},\vec{v})_0 \qquad \text{minden } \vec{v} \in V\text{-re}, \tag{7}$$

$$b(\vec{u},q) := (g,q)_0 \qquad \text{minden } q \in Q\text{-re}, \tag{8}$$

ahol teljesüljön $\int_\Omega g(x)\mathrm{d}x = 0$, és

$$a(\vec{u},\vec{v}) := (\vec{u},\vec{v})_1 = (-\Delta\vec{u},\vec{v})_0,$$
$$b(\vec{u},q) := -(\operatorname{div}\vec{u},q)_0.$$

Az átmenet (4)–(5)-ből (7)–(8)-ra úgy történik, hogy (4)-et skalárisan megszorozzuk a $\vec{v}$ tesztfüggvénnyel, majd parciálisan integráljuk (figyelembe véve a (6) peremfeltételeket). (5)-öt a q tesztfüggvénnyel szorozzuk meg és integráljuk.

A (7)–(8) variációs feladat megoldása, tehát a gyenge megoldás létezik, ha f, g olyanok, hogy $\vec{v} \to (\vec{f},\vec{v})_0$, ill. $q \to (g,q)_0$ folytonos lineáris funkcionálokat definiál V-n, ill. Q-n. Pl. ha $\vec{f} \in (L_2(\Omega))^d$, $g \in L_{2,0}(\Omega)$, ld. [8] vagy [22], és mivel a Lipschitz-folytonos Ω esetén teljesül az úgynevezett „inf-sup-feltétel", ld. 4. pont lent.

Ez a gyenge megoldás akkor egyértelmű, ha a p-vel kapcsolatos konstanst (ld. 1. pont) úgy határozzuk meg, hogy $\int_\Omega p\mathrm{d}x = 0$, azaz ha $p \in L_{2,0}(\Omega)$.

3. A Crouzeix–Velte-felbontás

A vektoranalízis ismert azonossága :

$$-\Delta = -\operatorname{grad}\operatorname{div} + \operatorname{rot}\operatorname{rot}, \tag{9}$$

vagyis (V-t a V' lineáris funkcionálok terébe képező operátorokkal)

$$A = B + C,$$
$$\text{ahol } A = -\Delta, \quad B = -\operatorname{grad}\operatorname{div}, \quad C = \operatorname{rot}\operatorname{rot},$$

és a homogén Dirichlet-féle peremfeltételek miatt igaz a következő ortogonális felbontás, [4], [23]:

$$\left(H_0^1(\Omega)\right)^d = V = V_0 \oplus V_1 \oplus V_\beta, \tag{10}$$

ahol az ortogonalitást a $\left(H_0^1(\Omega)\right)^d$ értelmében értjük, míg

$$V_0 := \ker\operatorname{div} = \{w \in V, \ \operatorname{div} w = 0\},$$
$$V_1 := \ker\operatorname{rot} = \{w \in V, \ \operatorname{rot} w = 0\}.$$

Itt a háromdimenziós esetben $\operatorname{rot} w$ a szokásos módon értendő, viszont a kétdimenziós esetben $\operatorname{rot} w$ jelöli a $\frac{\partial w_2}{\partial x_1} - \frac{\partial w_1}{\partial x_2}$ skalárt, amelyet néhol $\operatorname{curl} w$-ként is jelölnek.

Hogyan függ össze (9) és (10)?
Használjuk a 2. pont jelöléseit és integráljuk parciálisan a (9)-ből következő relációt

$$(\vec{u}, \vec{v})_1 = (-\Delta\vec{u}, \vec{v})_0 = (-\operatorname{grad}\operatorname{div}\vec{u}, \vec{v})_0 + (\operatorname{rot}\operatorname{rot}\vec{u}, \vec{v})_0.$$

A homogén peremfeltételek miatt ennek eredménye

$$(\vec{u}, \vec{v})_1 = (\operatorname{div}\vec{u}, \operatorname{div}\vec{v})_0 + (\operatorname{rot}\vec{u}, \operatorname{rot}\vec{v})_0. \tag{11}$$

Ha most $\vec{u} \in \ker\operatorname{div}$ és $\vec{v} \in \ker\operatorname{rot}$, akkor az egyenlőség jobb oldala nulla, vagyis $V_0 \perp V_1 = \ker\operatorname{div} \perp \ker\operatorname{rot}$ a V tér skalárszorzatának értelmében. Mind V_0, mind V_1 zárt alterek a V Hilbert-térben, és a $V_0 \oplus V_1$ altér ortogonális kiegészítő altere megint zárt altér V-ben – amit V_β-nak jelöltük.

4. A Schur-komplemens operátor és az inf-sup-feltétel

Legyen megint $A = -\Delta$, és mivel parciális integráció alapján grad adjungáltja $-\operatorname{div}$, írhatjuk a $B = -\operatorname{grad}\operatorname{div}$ operátort szorzatként mint $\tilde{B}^T\tilde{B}$, ahol $\tilde{B} = -\operatorname{div} : \ V \to Q$, $\tilde{B}^T = \operatorname{grad} : Q \to V'$. Akkor a (7)–(8) variációs feladatunk algebrai alakja:

$$\begin{pmatrix} A & \tilde{B}^T \\ \tilde{B} & 0 \end{pmatrix} \begin{pmatrix} \vec{u} \\ p \end{pmatrix} = \begin{pmatrix} \vec{f} \\ g \end{pmatrix}.$$

A Laplace-operátor esetünkben pozitív definit és szimmetrikus (a homogén Dirichlet-féle peremfeltételeknek köszönhetően), így egyértelműen invertálható:

$$\vec{u} = A^{-1}\left(\vec{f} - \tilde{B}^T p\right). \tag{12}$$

Ezt behelyettesítve:

$$g = \tilde{B}A^{-1}\left(\vec{f} - \tilde{B}^T p\right) = \tilde{B}A^{-1}\vec{f} - Sp, \tag{13}$$

ahol

$$S := \tilde{B}A^{-1}\tilde{B}^T \quad \left(= \operatorname{div}(\Delta)_0^{-1}\operatorname{grad}\right) \tag{14}$$

a Stokes-feladat Schur-komplemens operátora. (Itt a 0 index Δ^{-1}-nél a homogén peremfeltételre utal.) Ez az S operátor Q-ban működik és szinguláris, mert minden konstans nyomást a nullára képezi le, de a magteren kívül, $L_{2,0}(\Omega)$-ban egyértelműen és stabil módon invertálható, ha teljesül az „inf-sup"-feltétel:

Van olyan $\beta_0 = \beta_0(\Omega)$ pozitív konstans, amellyel igaz

$$\inf_{0\neq q\in L_{2,0}} \sup_{\vec{v}\in V} \frac{b(\vec{v},q)}{\|\vec{v}\|_1\|q\|_0} \geq \beta_0. \tag{15}$$

Ez a feltétel biztosított, mert Ω Lipschitz-folytonos tartomány.

Érvényes ugyanis Nečas tétele [12]: van olyan $c = c(\Omega)$ konstans, hogy a $-\operatorname{div}\vec{v} = q$ egyenlet adott $q \in L_{2,0}$ esetén olyan $\vec{v} \in V$-vel megoldható, amelyre igaz $\|\vec{v}\|_1 \leq c\|q\|_0$.

Ekkor

$$b(\vec{v},q) = \|q\|_0^2 \quad \text{és} \quad \|\vec{v}\|_1 \leq c\|q\|_0,$$

így

$$\sup_{\vec{v}\in V} \frac{b(\vec{v},q)}{\|\vec{v}\|_1} \geq \frac{1}{c}\|q\|_0, \quad \text{tehát} \quad \beta_0 \geq \frac{1}{c}.$$

Az inf-sup-feltétel a Stokes–Schur-komplemens operátor invertálhatóságát jelenti $L_{2,0}(\Omega)$-ban:

$$\left(\frac{b(\vec{v},q)}{\|\vec{v}\|_1}\right)^2 = \frac{(\tilde{B}\vec{v},q)_0^2}{(A\vec{v},\vec{v})_0} = \frac{(\tilde{B}A^{-1}\tilde{B}^T q,q)_0^2}{(\tilde{B}^T q, A^{-1}\tilde{B}^T q)_0} = (Sq,q)_0,$$

ha $\vec{v} = A^{-1}\tilde{B}^T q$ tetszőleges $0 \neq q \in L_{2,0}(\Omega)$ elemmel, így S pozitív sajátértékeinek alsó korlátja β_0^2 :

$$\beta_0^2 \leq \left(\frac{b(\vec{v},q)}{\|\vec{v}\|_1\|q\|_0}\right)^2 = \frac{(Sq,q)_0}{\|q\|_0^2}.$$

Felső korlátja 1, mivel homogén Dirichlet-féle peremfeltétel esetén (ld. (11)) igaz

$$\|\operatorname{div}\vec{v}\|_0^2 \leq \|\operatorname{div}\vec{v}\|_0^2 + \|\operatorname{rot}\vec{v}\|_0^2 = \|\vec{v}\|_1^2,$$

és innen következik

$$\frac{(Sq,q)_0}{\|q\|_0^2} = \left(\frac{(\operatorname{div}\vec{v},q)_0}{\|\vec{v}\|_1\|q\|_0}\right)^2 \leq \left(\frac{\|\operatorname{div}\vec{v}\|_0}{\|\vec{v}\|_1}\right)^2 \leq 1.$$

A Stokes–Schur-komplemens operátor spektrumáról, és azon belül a legkisebb pozitív $\lambda_1 = \beta_0^2$ sajátértékről, szerezhető részletesebb információ a [23], [9], [7], [11] cikkek alapján. Innen tudható pl., hogy (mindig homogén Dirichlet-féle peremfeltétel esetén) a sajátértékek $[0,1]$-ből valók, hogy a kétdimenziós esetben a spektrum szimmetrikus $1/2$-re (ha λ sajátérték, akkor $1-\lambda$ is az, sőt, ha $\lambda \in (0,1)$, akkor ez egyenlő multiplicitás mellett érvényes). Maga az $1/2$ végtelen multiplicitású sajátérték, mint ahogyan 1 is ilyen. Továbbá, $\lambda_1 \leq \frac{1}{2} - \frac{\sin\omega}{2\omega}$, ahol ω a legkisebb belső szög $\leq \pi$ a tartomány peremén. A spektrum további tulajdonságait vizsgálják a tartomány peremének függvényében pl. a [26] és [27] cikkek. Amikor a peremen egy π-től különböző belső szög fordul elő, a spektrumnak folytonos része is van.

A háromdimenziós esetnek megfelelő S operátor spektrumáról lényegesen kevesebbet ismerünk, ld. pl. [5] és [23].

Befejezésül arra mutatunk rá, hogy az egyenletek diszkretizációja után a stabilitáshoz fontos a megfelelő *diszkrét* inf-sup feltétel. Ekkor (15)-ben a megfelelő diszkrét terek állnak, és a β_0 pozitív konstans h-tól független, ill. a megfelelő diszkrét S operátor legkisebb pozitív sajátértéke egy h-tól független konstanssal el van választva a nullától. Ezt a diszkrét inf-sup feltételt a diszkrét nyomás- és sebességtér alkalmas választásával lehet biztosítani, a diszkrét feltétel bizonyítása nem egyszerű és a mindenkori alkalmas térkombinációtól függ [2].

5. Az algebrai és a diszkrét Crouzeix–Velte-felbontás

Legyenek most $A, B, C \in \mathbb{R}^{n\times n}$ szimmetrikus és pozitív szemidefinit mátrixok, $A = B + C$, de A pozitív definit, úgy hogy (Au, v) skalárszorzat.

Ekkor

$$(Au, v) = (Bu, v) + (Cu, v).$$

Ebből azt látjuk, úgy, mint fent (11)-ből (legyen $u \in \ker B$, $v \in \ker C$ stb.), hogy érvényes (10) algebrai hasonmása, a következő **„algebrai Crouzeix–Velte-felbontás"**:

$$\mathbb{R}^n = \ker B \oplus \ker C \oplus W, \tag{16}$$

ahol az ortogonalitást $(A\cdot, \cdot)$ értelmében értjük.

Ez az ortogonális felbontás jellemezhető sajátértékekkel. Legyen ugyanis

$$\lambda Au = Bu \quad (\text{ill. } \mu Av = Cv). \tag{17}$$

Ekkor egyrészt $A = B + C$ miatt $\mu_i = 1 - \lambda_i$, másrészt

$$\ker B = \operatorname{span}\left(u^{(i)}, \lambda_i = 0\right),$$
$$\ker C = \operatorname{span}\left(u^{(i)}, \lambda_i = 1\right),$$
$$W = \operatorname{span}\left(u^{(i)}, \lambda_i \in (0,1)\right).$$

Összehasonlítva (10)-et és (16)-ot, azt látjuk, hogy a sebességi térben a 0 sajátértékek a divergenciamentes sebességeknek felelnek meg, az 1 sajátértékek a rotációmentes sebességeknek. Tehát ami itt az algebrai esetben szerepel, az az analitikus esetben is érvényes: a (10)-beli alterek indexei erre a sajátértékekkel való jellemzésre utalnak.

A legérdekesebb a harmadik altér, V_β, ill. W, amelynek függvényei se nem divergencia-, se nem rotációmentesek. A [15] dolgozatban egyebek mellett több feltételt adtunk, hogy ez a harmadik altér mikor nem triviális.

A (10) Crouzeix–Velte-felbontásnál ismertebb a Helmholtz–Weyl-felbontás [25], amelynek alakja

$$L_2 = \ker \operatorname{div} \oplus \ker \operatorname{rot} \oplus H.$$

Itt viszont a harmadik tér függvényei divergencia- és rotációmentesek, azaz harmonikusak. A közös tulajdonság az, hogy mindkét harmadik tér „peremeffektusokkal" kapcsolatos: egy harmonikus függvény már peremértékeivel definiált, és a V_β tér elemei biharmonikusok és a Neumann-peremértékeivel definiáltak [15] (míg Dirichlet-peremértékei homogének $V_\beta \subset V$-nek megfelelően).

Amennyiben a fenti B mátrix felbontható $B = \tilde{B}^T\tilde{B}$ alakban, akkor a $p = \tilde{B}u$ transzformációval (17)-ből nyerhetjük az $S := \tilde{B}A^{-1}\tilde{B}^T$ algebrai Stokes–Schur-komplemens operátorra (vö. (14)-gyel) vonatkozó

$$\lambda p = Sp$$

sajátérték feladatot, és ezalatt a transzformáció alatt csak a nulla sajátérték multiplicitása változhat – maguk a sajátértékek nem. Ez a (4)–(6) feladat bizonyos megfelelő diszkretizációihoz tartozó S operátorra azt jelenti, hogy a nyomás végeselem térében az 1 sajátérték multiplicitása magas, a nulla sajátérték multiplicitása 1. Közben az 1, ill. 0 sajátértékhez (17) révén tartozó sebességek a végeselem térben rotációmentesek, ill. divergenciamentesek.

Az algebrai Crouzeix–Velte-felbontás alapján meg lehet mutatni, hogy valóban vannak olyan differencia [13], [6], véges térfogat [20], [21] és végeselem approximációi [6], [15], [18], [19], [1] a Stokes-feladatnak, ahol az összes feltétel ($A = B + C$, A a $-\Delta$ operátornak megfelelő mátrix, tehát szimmetrikus és pozitív definit stb.) érvényes, így **diszkrét Crouzeix–Velte-felbontás**ról is lehet beszélni.

Emellett a kétdimenziós esetnek olyan bevált diszkretizációi is vannak, mint [3], amelyeknek nincsen olyan 3 alteres felbontásuk, amelyben a rotáció-, ill. divergenciamentes sebességek 1-1 alteret alkotnának, közben a hozzátartozó S mátrix spektruma nem $[0,1]$-ben van – mint az analitikus esetben –, hanem $[0,2)$-ben (sőt: $h \to 0$ esetén vannak 2-höz tartó sajátértékek) és magas multiplicitású sajátérték nincsen.

6. Iterációs módszerek a diszkrét Stokes-feladat megoldására

Az algebrai Crouzeix–Velte-felbontás megléte nemcsak szép matematikailag, hanem a numerikus megoldás során előnyös, mint ahogyan igazoltuk [14]-ben az

Uzawa, Arrow–Hurwitz és a konjugált gradiens módszerekre: gyorsabb konvergenciára ad lehetőséget és az Uzawa-, valamint az Arrow–Hurwitz-algoritmusok esetén optimális iterációs paraméterekre. (Speciális esetben, az egységnégyzetben megfogalmazott Stokes-feladat „staggered grid" approximációja esetén, ilyen lehetőséget már [10]-ben is vizsgálták.)

Példaként az Uzawa-iterációt említjük, amely a (12)–(14) egyenletek diszkretizált verziójából jön létre:

$$\begin{aligned} p_0 \text{ adott, } \quad k &= 0,1,\ldots\text{-ra:} \\ \vec{u}_k &= A^{-1}(\vec{f} - \tilde{B}^T p_k), \\ p_{k+1} &= p_k + \omega(\tilde{B}\vec{u}_k - g), \end{aligned}$$

ahol ω az iterációs paraméter.

Ez az ún. egyszerű iteráció a (diszkrét) nyomás-térbeli (13) egyenlet megoldására, ugyanis a második sort a harmadikba behelyettesítve következik

$$p_{k+1} = p_k + \omega(\tilde{B}A^{-1}(\vec{f} - \tilde{B}^T p_k) - g) = (I - \omega S)p_k + \omega(\tilde{B}A^{-1}\vec{f} - g).$$

Innen kapjuk a p_k közbülső nyomás $e_k := p_k - p$ hibájára (legyen p a diszkretizált (13) egyenlet pontos megoldása), hogy

$$e_{k+1} = (I - \omega S)e_k.$$

Ez a hibaegyenlet azonnal azt mutatja, hogy a diszkrét $\ker S$ térrel nem kell foglalkoznunk. Ez az altér remélhetőleg (és megfelelő diszkretizáció mellett valóban) egydimenziós, mint az analitikus feladat esetén a konstans nyomások altere, és e_k-nak ezen altérbeli része az iteráció alatt nem változik, így p_k-nak ezen része sem változik. Akkor ez a rész p_0-tól függ, amely tetszőlegesen választható – ami annak felel meg, hogy a feladat szinguláris.

A gyors konvergencia most a következőkkel kapcsolatos:

1. Egyetlen $\omega = 1$-gyel végrehajtott legelső iterációval eltüntethető az összes, a magas multiplicitású 1 sajátértékhez tartozó függvényrész a hibából, hiszen ha v bármely az 1 sajátértékhez tartozó sajátvektora S-nek, akkor érvényes $(I - S)v = 0$. Vagyis (a sebességek szintjén) eltűnt az egész diszkrét $\ker \operatorname{rot}$ altér a hibából.
2. Ezután a maradékhiba már W-ben van, így csak (alacsony dimenziójú) peremeffektusokkal kapcsolatos: W dimenziója a perempontok számával arányos.
3. A W-re szűkített operátor spektruma h-tól majdnem függetlenül igen keskeny. Itt most fontos a spektrum határainak tudása, és ha a W-re szűkített spektrum eleget tesz a

$$0 < m \leq \lambda^h_{\min} \leq \lambda^h_{\max} \leq M$$

egyenlőtlenségnek, akkor m és M ismeretében az optimális iterációs paraméter szokásos módon [16] $\omega = \frac{2}{m+M}$-ből számítható.

4. Ha a kétdimenziós esetben az előzőkhöz hozzájön, hogy a diszkretizáció is egy 1/2-re szimmetrikus spektrumot eredményez (a „konform" végeselem esetben ez áll fenn – sőt ebben az esetben a tulajdonság független a tartománytól és triangulációjától [15]), akkor $\lambda^h_{\max} = 1 - \lambda^h_{\min}$, tehát ebben az esetben az optimális iterációs paraméter $\omega = \frac{2}{\lambda^h_{\min}+\lambda^h_{\max}} = 2$, függetlenül a tartománytól és a triangulációtól!

A kétdimenziós Stokes-feladat differencia és véges térfogat approximációi rendszerint az S spektrumának szimmetriáját 1/2-re nem adják vissza. Ekkor tanácsolható, hogy durva rácson számítsuk ki a $\lambda^h_{\min}$, $\lambda^h_{\max}$ értékeit, majd finom rácson használjuk, mert h-val csak lassan változnak.

7. Néhány számítási eredmény

Az alábbi számszerű eredményeket [14]-ből idézzük. Elsőnek tekintsük a Stokes-feladatot az egységnégyzetben, homogén elsőfajú peremfeltételek mellett, felosztva ezt a négyzetet $(N-1)^2$ kis négyzetre, mindegyik $h = 1/(N-1)$ hosszú oldalokkal. A feladatot diszkretizáltuk az ismert eltolt rácsrendszerű differencia approximáció segítségével, ld. [6], [13] vagy [17], 342. o. Ekkor az S spektrumának szimmetriája 1/2-re nem áll fenn, és például $N = 256$ esetén az optimális iterációs paraméter 2 helyett 1.84 volt.

Továbbá, létrehozva az adott diszkretizációhoz egy algebrai Stokes-feladatot ismert egzakt megoldással, a fenti elméletnek megfelelően egy Uzawa-lépést tettünk ω = 1-gyel, majd ezután egyszer a spektrumból kiszámított ω_{opt} paraméterrel folytattuk, egyszer viszont $\omega = 2$-vel. Ennek során számoltuk azon lépések $it(\omega)$ számát (beleértve az ω = 1-féle lépést), amely elegendő ahhoz, hogy a kiindulási nyomás-hibát az euklideszi normában a 10^{-10}-ed részére csökkentse.

N	4	8	16	32	64	128	256
$it\left(\tau^h_{\text{opt}}\right)$	14	20	26	31	35	39	41
q	0.1868	0.3153	0.4098	0.4728	0.5258	0.5458	0.5646
$it(2)$	62	58	55	56	55	55	53
q	0.6873	0.6722	0.6555	0.6603	0.6576	0.6540	0.6471

A táblázatban a q konvergencia rátát is mutatjuk. Ezt a nyomás $e_0 := \|e^{(0)}\|$ kezdeti és $e_m := \|e^{(m)}\|$ végső hibájának euklideszi normájából számítottuk ki: $q := (e_m/e_0)^{1/m}$, ahol $m = it(\omega)$.

Ezekből az eredményekből kirajzolódik, hogy $N \to \infty$ esetén q tart egy 0.6 körüli értékhez – és nem 1-hez. Ilyen értelemben a konvergencia h-tól független.

Következőnek a [24]-féle konjugált gradiens módszert alkalmazzuk a diszkretizált (13) egyenletek megoldására.

Különböző diszkretizációkat vizsgálva kísérletileg kiderül, hogy olyan diszkretizáció egyenleteit hatékonyabban meg lehet oldani, amelyeknél a diszkrét Crouzeix–Velte-felbontás létezik.

Az alábbi táblázatban SG jelenti az eltolt rácsrendszer (staggered grid) approximációját, Q1 a $Q_1 - Q_1$(macro)-végeselemeket [2] és CR az ismert Crouzeix–Raviart-elemeket [3]. Itt is $N = 1 + 1/h$, emögött mutatjuk a megfelelő diszkrét nyomás-tér dimenzióját (és igyekeztünk ezeknek lehetőleg közeli értékeket adni N választásával). A kilépési kritérium most az volt, hogy a maradékvektor végső és kiindulási euklideszi normáinak hányadosa 10^{-10} legyen, és it az ehhez szükséges lépésszám.

SG	$N(\dim P_h)$	7(36)	12(121)	24(529)	46(2025)
	it	10	12	14	15
	q	0.0728	0.1307	0.1706	0.2043
Q1	$N(\dim P_h)$	13(36)	23(121)	47(529)	91(2025)
	it	16	18	19	20
	q	0.2292	0.2602	0.2930	0.3061
CR	$N(\dim P_h)$	5(32)	9(128)	17(512)	33(2048)
	it	19	23	26	27
	q	0.2811	0.3647	0.4038	0.4225

Megjegyezzük, hogy a három módszer közül csak SG rendelkezik diszkrét Crouzeix–Velte-felbontással.

8. További feladatok

A homogén Dirichlet-feltétel mellett további peremfeltételekre sikerült a Crouzeix–Velte-felbontás meglétét úgy a Stokes-rendszer analitikus esetében, mint diszkretizációja után kimutatni, ld. [18]: pl. ha kétdimenziós esetben a tartomány téglalap, és a vízszintes oldalakon Dirichlet-feltétel van, a függőleges oldalakon periodikus feltétel. Általános tartomány esetén olyan nemstandard, de hasznos peremfeltétel is alkalmas a Crouzeix–Velte-felbontás megléte szempontjából, ahol a sebesség normálkomponense és rotációja nulla a peremen.

Az az eset, amikor a peremnek egyik részén Dirichlet-, a másik részén Neumann-feltétel adott, gyakori az áramlási feladatoknál, de nem vezet a három alteres Crouzeix–Velte-felbontásra. Ilyenkor viszont lehet tudni, hogy a spektrum $[0, d]$-ben van $(d = 2, 3)$, és $\omega = d/2$ alkalmas iterációs paraméter.

Befejezésül megemlítjük, hogy van egy további fontos példa a Crouzeix–Velte-felbontás hasznosíthatóságára: a linearizált rugalmasságtani egyenletek [18]:

$$-(2\mu + \lambda)\,\mathrm{grad}\,\mathrm{div}\,\vec{u} + \mu\,\mathrm{rot}\,\mathrm{rot}\,\vec{u} = \vec{f}, \quad x \in \Omega,$$
$$\vec{u} = 0, \quad x \in \Gamma.$$

Amennyiben

$$\mu > 0 \text{ és } \lambda + \mu \geq 0,$$

akkor a Crouzeix–Velte-felbontás létezik. Itt

$$\sqrt{\mu}\,\mathrm{rot} =: \tilde{C}, \qquad \sqrt{2\mu+\lambda}\,\mathrm{div} =: \tilde{B},$$
$$A = B + C, \qquad B = \tilde{B}^T\tilde{B}, \qquad C = \tilde{C}^T\tilde{C}.$$

Ekkor $a(\vec{u},\vec{v}) = (A\vec{u},\vec{u})$ szimmetrikus és pozitív definit (érvényes a Korn-féle egyenlőtlenség) :

$$a(\vec{u},\vec{u}) \geq \mu\left\{(\mathrm{div}\,\vec{u},\mathrm{div}\,\vec{u})_0 + (\mathrm{rot}\,\vec{u},\mathrm{rot}\,\vec{u})_0\right\}.$$

Hivatkozások

[1] Baran Á., Stoyan G.: *Gauss–Legendre elements: a stable, higher order non-conforming finite element family.* Computing **79**,1 (2007), 1–21.

[2] Brezzi, F., Fortin, M.: *Mixed and Hybrid Finite Element Methods.* New York: Springer 1991.

[3] Crouzeix, M., Raviart, P.–A.: *Conforming and nonconforming finite element methods for solving the stationary Stokes equations.* R.A.I.R.O. - Informatique Théoretique et Applications, R-3 (1973), 33–76.

[4] Crouzeix M.: *Étude d'une méthode de linéarisation. Résolution numérique des équations Stokes stationnaires.* In: Cahier de l'INRIA **12**, (1974), 139–244.

[5] Crouzeix, M.: *On an operator related to the convergence of Uzawa's algorithm for the Stokes equation.* In: Computational Science for the 21st Century (J. Périaux et al., eds.). New York: Wiley 1997, 242–249.

[6] Dobrowolski M., Stoyan G.: *Algebraic and discrete Velte decompositions.* BIT **41**, No. 3 (2001), 465–479.

[7] Friedrichs K.O.: *On certain inequalities and characteristic value problems for analytic functions of two variables.* Trans. Amer. Math. Soc. **41** (1937), 321–364.

[8] Girault V., Raviart P.-A.: *Finite Element Methods for Navier-Stokes Equations.* Springer, Berlin 1986.

[9] Horgan C.O., Payne L.E.: *On inequalities of Korn, Friedrichs and Babuška-Aziz.* Arch. Rational Mech. Anal. **82** (1983), 165–179.

[10] Kobelkov G.M., Orlova N.B.: *A fast method for solving the Stokes problem.* Vestnik Moskov. Univ. Ser. XV, Vychisl. Mat. Kibernet. **4** (1989), 39–45 (in Russian).

[11] Mikhlin S.G.: *The spectrum of the operator pencil of elasticity theory.* Uspekhi Mat. Nauk **28**, No. 3 (171) (1973), 43–82 (in Russian).

[12] Nečas J.: *Equations aux Dérivées Partielles.* Presses de l'Université de Montreal 1965.

[13] STOYAN G.: *Towards discrete Velte decompositions and narrow bounds for inf-sup constants.* CAMWA **38** (1999), 243–261.

[14] STOYAN G.: *Iterative Stokes solvers in the harmonic Velte subspace.* Computing **67** (2000), 13–33.

[15] STOYAN G.: $-\Delta$ = *-grad div + rot rot for matrices, with application to the finite element solution of the Stokes problem.* East-West Journal of Numerical Mathematics 8, No. 4 (2000), 323–340.

[16] STOYAN G., TAKÓ G.: *Numerikus módszerek I*, 2. kiadás Typotex, Budapest 2002. (A 3. kiadás interneten olvasható, ld. a http://numanal.inf.elte.hu/∼ stoyan honlapon „books" alatt.)

[17] STOYAN G., TAKÓ G.: *Numerikus módszerek III*, Typotex, Budapest 1997.

[18] STOYAN G., STRAUBER GY., BARAN Á.: *Generalizations to discrete and analytical Crouzeix-Velte decompositions.* Numer. Lin. Algebra **11** (2004), 565-590.

[19] STOYAN G., BARAN Á.: *Crouzeix-Velte decompositions for higher-order finite elements.* CAMWA **51** (2005), 967–986.

[20] STRAUBER GY.: *Discrete Crouzeix-Velte decompositions on nonequidistant rectangular grids.* Annales Univ. Budapest **44** (2002), 63–82.

[21] STRAUBER GY.: *Discrete Crouzeix-Velte decomposition for the disk domain.* Miskolc Mathematical Notes **6**,1 (2005), 129-141.

[22] VARNHORN W.: *The Stokes Equations.* Mathematical Research **76**, Akademie-Verlag Berlin 1994.

[23] VELTE W.: *On optimal constants in some inequalities.* Lecture Notes in Math. 1431, 158–168. Springer, Berlin 1990.

[24] VERFÜRTH, R.: *A combined conjugate gradient–multigrid algorithm for the numerical solution of the Stokes problem.* IMA J. Numer. Anal. **4**, 441–455 (1984).

[25] WEYL H.: *The method of orthogonal projection in potential theory.* Duke Math. J. **7** (1940), 411–444.

[26] ZSUPPAN S.: *On the spectrum of the Schur complement of the Stokes operator via conformal mapping.* Methods and Applications of Analysis **11**,1 (2004), 133–154.

[27] ZSUPPÁN S.: *On connections between the Stokes-Schur and the Friedrichs operator, with applications to the inf-sup problem.* Annales Univ. Budapest **48** (2005), 151–171.

STOYAN GISBERT
Eötvös Lóránd Tudományegyetem
1117 Budapest, Pázmány Péter sétány 1/c
stoyan@numanal.inf.elte.hu

THE STOKES PROBLEM AND THE CROUZEIX-VELTE DECOMPOSITION

STOYAN GISBERT

In this paper a survey of the author's and his disciples' results is presented on the research on the discretizations for solving the first-type boundary value problem for the Stokes equations so that the discretizations preserve the existence of a special 3-subspace orthogonal decomposition. In this way not only a characteristic property of the analytic case is preserved, but also the efficiency of the iterative solution methods for the discretized equations is significantly improved. This technique is transmissible to other problems.

Alkalmazott Matematikai Lapok **26** (2009), 193-206.

EGY MAGAS RENDŰ NEMKONFORM VÉGESELEM CSALÁD A KÉTDIMENZIÓS STOKES-FELADAT MEGOLDÁSÁRA

BARAN ÁGNES

A cikkben a kétdimenziós Stokes-feladat numerikus megoldása kapcsán egy nemkonform, háromszöges végeselem családdal foglalkozunk. Hasonlóan a Scott és Vogelius által definiált konform elempárhoz a nyomást és a sebesség koordinátafüggvényeit itt is háromszögenként $(k-1)$-edfokú és k-adfokú polinomokkal approximáljuk. A diszkrét sebességek esetén – eltérően a Scott–Vogelius-elemtől – a folytonosságot a szomszédos háromszögek közös oldalain csak bizonyos pontokban követeljük meg. A végeselem pár tetszőleges k rend esetén definiált és ismert alacsony rendű ($k = 1, 2, 3$) elemek általánosítása. Megmutatjuk, hogy páros k esetén az elem a Scott–Vogelius-elemből származtatható, annak sebességi terét háromszögenként egy nemkonform buborékfüggvénnyel bővítve. A buborékfüggvény megszünteti a Scott–Vogelius-elemekkel való diszkrét megoldás során felmerülő esetleges algebrai szingularitást (az „energiamentes" diszkrét nyomásfüggvények jelenlétét). Belátjuk, hogy páros k esetén az elempár stabil.

1. Bevezetés

Legyen $\Omega \subset \mathbb{R}^2$ egy poligonális tartomány $\Gamma = \partial\Omega$ határral. A viszkózus, összenyomhatatlan folyadékok stacionárius áramlását leíró egyenletrendszer:

$$\begin{aligned} -\nu \cdot \Delta\vec{u} + \frac{1}{\rho} \cdot \mathrm{grad}P &= \vec{f} \quad \Omega\text{-n}, \\ \mathrm{div}\vec{u} &= 0 \quad \Omega\text{-n}, \\ \vec{u}|_\Gamma &= 0, \end{aligned}$$

ahol ν a kinematikus viszkozitás, $\vec{u}$ a sebességvektor, ρ a sűrűség, P a nyomás, $\rho\vec{f}$ a külső erők vektora. Feltételezve, hogy ρ pozitív konstans, bevezetve a $p = P/\rho$ jelölést és a ν viszkozitást 1-nek választva a fenti egyenletrendszer

$$\begin{aligned} -\Delta\vec{u} + \mathrm{grad}p &= \vec{f} \quad \Omega\text{-n}, \\ \mathrm{div}\vec{u} &= 0 \quad \Omega\text{-n}, \\ \vec{u}|_\Gamma &= 0, \end{aligned} \tag{1}$$

alakba írható, (1)-et stacionárius Stokes-feladatnak nevezzük. Az egyenletekből p csak egy additív konstanstól eltekintve határozható meg egyértelműen. Az inhomogén Dirichlet-feltétel az (1)-ben szereplő homogén peremfeltételre visszavezethető.

Jelen cikkben a Stokes-feladat végeselem megoldásával kapcsolatban vizsgálunk magas rendű háromszöges elemeket. A magas rendű véges elemek hasznáról ld. például [7], [13].

Egy adott végeselem diszkretizáció kapcsán mindig felmerül az a kérdés, hogy az egyértelmű, stabil megoldás létezését biztosító inf-sup feltétel teljesül-e (ld. a cikk 2. pontját).

Scott és Vogelius [10] a konform $\mathbb{P}_k/\mathbb{P}_{k-1}$ végeselem párt vizsgálta (háromszögenként k-adrendű polinomok a sebesség, és $k-1$-edrendű polinomok a nyomás közelítésére), ahol a diszkrét nyomás függvényekről nem feltételezzük, hogy folytonosak a háromszögek találkozásánál. $k \geq 4$ esetén az elem stabil, de csak egy a rácsra vonatkozó feltétel teljesülése esetén.

Egy másik esetleges probléma a konform $\mathbb{P}_k/\mathbb{P}_{k-1}$ elemek használatánál, hogy a diszkrét gradiens operátor nulltere nagyobb lehet, mint az eredeti problémában a gradiens operátor nulltere, amely csak a konstans függvényeket tartalmazza. Ez azt jelenti, hogy amíg folytonos esetben a nyomás egy additív konstantól eltekintve egyértelműen meghatározható, addig a végeselem diszkretizáció után kapott lineáris egyenletrendszernek többdimenziós nulltere van.

A rácstól független stabilitás kapcsán kerülnek előtérbe a nemkonform elemek: itt a sebességet approximáló háromszögenként definiált k-adrendű polinomok folytonosságát a szomszédos háromszögek közös oldalain csak bizonyos pontokban követeljük meg. A 3. fejezetben egy tetszőleges k rend esetén definiált, nemkonform végeselem párt írunk le, amely minden páros k esetén stabil, a rácsra vonatkozó feltétel nélkül. Páros rend esetén a diszkrét sebesség tér a k-adrendű konform elem sebesség terének bővítése: annak bázisához háromszögenként egy k-adfokú polinomot, egy úgy nevezett nemkonform buborék függvényt adunk. $k = 2$ esetén az elem megegyezik az ismert Fortin–Soulie-elemmel [6], a $k = 4, 6$ esetekben pedig a [4]-ben vizsgált nemkonform elemekkel. Amíg [4]-ben a sebességi tér leírásánál használt buborékfüggvényt csak a $k = 4, 6$ esetben sikerült képlettel leírni, addig a 3. fejezetben tetszőleges páros k esetén érvényes formulát adunk. [4]-ben a szerzők belátták a negyed-, és hatodrendű elem stabilitását, de a szokásos $b(\cdot, \cdot)$ bilineáris formát (ld. 2. pont) kiegészítették egy stabilizáló taggal.

Matthies és Tobiska [9] egy tetszőleges rend esetén definiált stabil, nemkonform végeselem családot írnak le, de a k-adrendű konform sebességi teret egy háromszögenként $(k+1)$-edrendű polinommal bővítik.

Belátjuk, hogy a nemkonform buborék függvény hatására a diszkrét gradiens operátor nulltere egydimenziós lesz, függetlenül a rácstól.

2. Jelölések, alapfogalmak

Az (1) feladat gyenge megfogalmazásához vezessük be az alábbi jelöléseket. Legyen $L^2(\Omega)$ a négyzetesen integrálható függvények tere, továbbá

$$L_0^2(\Omega) = \left\{ p \in L^2(\Omega) : \int_\Omega p \mathrm{d}x = 0 \right\}.$$

A p függvényt – hogy az ne csak egy additív konstans erejéig legyen egyértelmű – az $L_0^2(\Omega)$ térben fogjuk keresni. A négyzetesen integrálható gradienssel rendelkező $L^2(\Omega)$-beli függvények Szoboljev-terét jelölje $(H^1(\Omega))^2$, és $(H_0^1(\Omega))^2$ legyen azon $(H^1(\Omega))^2$-beli függvények tere, melyek nyoma eltűnik Γ-n. Ezek után az (1) feladat gyenge megfogalmazása a következő: olyan $\vec{u} \in (H_0^1(\Omega))^2$ és $p \in L_0^2(\Omega)$ függvényeket keresünk, melyekre

$$\begin{aligned} a(\vec{u},\vec{v}) + b(\vec{v},p) &= (\vec{f},\vec{v}) \qquad \forall \vec{v} \in (H_0^1(\Omega))^2, \\ b(\vec{u},q) &= 0 \qquad\qquad \forall q \in L_0^2(\Omega) \end{aligned} \tag{2}$$

teljesül, ahol $(\cdot,\cdot)$ jelöli az $L^2(\Omega)$ és az $(L^2(\Omega))^2$ tér szokásos belső szorzatát is, továbbá

$$a(\vec{u},\vec{v}) = \int_\Omega \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\partial u_i}{\partial x_j}\frac{\partial v_i}{\partial x_j}\mathrm{d}x, \tag{3}$$

$$b(\vec{v},p) = -(\mathrm{div}\vec{v},p). \tag{4}$$

A feladat egyértelműen megoldható, ha $\vec{f} \in (L_2(\Omega))^2$.

Legyen $\mathcal{T}_h$ az Ω tartomány egy triangularizációja, továbbá jelölje E a triangularizáció éleinek halmazát.

A (2) feladat végeselem megoldása során a sebességkomponenseket és a nyomást háromszögenként adott fokszámú polinomokkal közelítjük, jelölje $V_h(\Omega)$, ill. $P_h(\Omega) \subset L^2(\Omega)$ a diszkrét sebesség, ill. nyomás tereket. Ha $V_h(\Omega) \subset (H_0^1(\Omega))^2$ teljesül konform, ellenkező esetben nemkonform approximációról beszélünk.

Ekkor a (2) egyenleteknek megfelelő diszkrét feladat: olyan $\vec{u}_h \in V_h(\Omega)$ és $p_h \in P_h(\Omega)$ függvényeket keresünk, melyekre

$$\begin{aligned} a(\vec{u}_h,\vec{v}_h) + b(\vec{v}_h,p_h) &= (\vec{f},\vec{v}_h) \qquad \forall \vec{v}_h \in V_h(\Omega), \\ b(\vec{u}_h,q_h) &= 0 \qquad\qquad \forall q_h \in P_h(\Omega) \cap L_0^2(\Omega) \end{aligned} \tag{5}$$

teljesül, ahol nemkonform esetben az $a(\cdot,\cdot)$ és $b(\cdot,\cdot)$ funkcionálokat a következő módon definiáljuk:

$$a(\vec{u},\vec{v}) = \sum_{\Delta\in\mathcal{T}_h} \int_\Delta \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\partial u_i}{\partial x_j}\frac{\partial v_i}{\partial x_j}\mathrm{d}x, \tag{6}$$

$$b(\vec{v},p) = -\sum_{\Delta\in\mathcal{T}_h} \int_\Delta p\mathrm{div}\vec{v}\mathrm{d}x. \tag{7}$$

Konform esetben (6)–(7) ugyanazokat a funkcionálokat definiálja, mint (3)–(4). Legyen

$$N_{V_h(\Omega)} := \{p_h \in P_h(\Omega) : b(\vec{v}_h, p_h) = 0 \quad \forall \vec{v}_h \in V_h(\Omega)\}$$

a diszkrét gradiens operátor nulltere.

Ahhoz, hogy az (5) feladat egyértelműen megoldható legyen a $V_h(\Omega)$, $(P_h(\Omega)\backslash N_{V_h(\Omega)})$ terekben szükséges az úgynevezett diszkrét inf-sup feltétel teljesülése, azaz létezzen olyan $\beta_h > 0$ konstans, hogy

$$\sup_{\vec{v}_h \in V_h} \frac{b(\vec{v}_h, q_h)}{|\vec{v}_h|_1} \geq \beta_h \|q_h\|_{L^2(\Omega)} \qquad \forall q_h \in P_h(\Omega)\backslash N_{V_h(\Omega)},$$

ahol $|\vec{v}_h|_1 = (a(\vec{v}_h, \vec{v}_h))^{1/2}$. Ha $\beta_h \geq \beta > 0$, ahol β független h-tól, akkor a végeselem megoldás stabil (ld. [3]), azaz

$$\|\vec{u}_h\|_1 \leq C_1 \cdot \|\vec{f}\|_{L^2(\Omega)}, \qquad \|p_h\|_0 \leq C_2 \cdot \|\vec{f}\|_{L^2(\Omega)}.$$

Ebben az esetben a $V_h(\Omega)$, $P_h(\Omega)$ végeselem párt stabilnak (vagy inf-sup stabilnak) nevezzük.

Ha az inf-sup feltétel nem teljesül, akkor tipikusan a sebességek konvergálnak, a nyomás viszont nem.

Ha $|\cdot|_1$ normát definiál a $V_h(\Omega)$ téren, akkor az inf-sup feltétel elegendő is az egyértelmű megoldás létezéséhez.

3. A Gauss–Legendre-elemek

Jelölje $\mathbb{P}_k(\Delta)$ a Δ háromszögön definiált legfeljebb k-adfokú polinomok terét, és tegyük fel, hogy a $\mathcal{T}_h$ triangularizácó reguláris, azaz létezik olyan h-tól független κ konstans, hogy

$$h_\Delta \leq \kappa \rho_\Delta \qquad \forall \Delta \in \mathcal{T}_h,$$

ahol h_Δ a Δ háromszög átmérője, ρ_Δ pedig a Δ háromszögbe írható körök sugarainak maximuma.

Vizsgálataink kiindulópontja Scott és Vogelius [10] cikke, melyben a konform, háromszöges $\mathbb{P}_k/\mathbb{P}_{k-1}$ elemekkel foglalkoztak: itt a sebességet háromszögenként k-adrendű, a háromszögek között folytonos polinomokkal, míg a nyomást háromszögenként $(k-1)$-edrendű polinomokkal approximálták (a diszkrét nyomás folytonossága nem feltétel). A megfelelő diszkrét terek:

$$P_h(\Omega) = \left\{p_h \in L^2(\Omega) : p_h|_\Delta \in \mathbb{P}_{k-1}(\Delta), \Delta \in \mathcal{T}_h\right\}, \tag{8}$$

$$V_h(\Omega) = \left\{\vec{v}_h \in (H_0^1(\Omega))^2 : \vec{v}_h|_\Delta \in (\mathbb{P}_k(\Delta))^2, \Delta \in \mathcal{T}_h\right\}. \tag{9}$$

$k \geq 4$ esetén a (8)–(9) végeselem pár stabilitása és diszkrét gradiens operátor nullterének dimenziója attól függ, hogy a rács tartalmaz-e közel-szinguláris, ill. szinguláris pontokat.

3.1. Definíció. Legyen x_0 a $\mathcal{T}_h$ triangularizáció egy rácspontja. Jelölje Δ_i, $i = 1, \ldots, n$, a triangularizáció azon háromszögeit, melyeknek x_0 csúcspontja és legyen Θ_i a Δ_i háromszög x_0-beli szöge. Tegyük fel, hogy a háromszögek sorszámozása olyan, hogy a Δ_i és Δ_{i+1} háromszögeknek van közös oldala minden $i = 1, \ldots, n-1$ esetén. Ekkor az x_0 pontot szinguláris pontnak nevezzük, ha $\Theta_i + \Theta_{i+1} = \pi$, $i = 1, \ldots, n-1$, teljesül. Az x_0 szinguláris pont belső szinguláris pont, ha $x_0 \in \Omega \backslash \Gamma$ (ekkor $n = 4$), és perem szinguláris pont, ha $x_0 \in \Gamma$ (ekkor $n \leq 4$).

Szemléletesen, szinguláris pontnak nevezünk egy rácspontot, ha a triangularizáció azon élei, melyek tartalmazzák az adott rácspontot két (egymást a csúcsban metsző) egyenesen fekszenek. A 3. ábrán egy belső szinguláris pont látható, míg a perem szinguláris pont 4 típusát az 1. és 2. ábrán mutatjuk meg. A szinguláris pontot mindhárom ábrán S_0 jelöli, a tartomány peremét vastag vonallal jelöltük.

A [10]-ben leírtak alapján belátható, hogy:

3.1. ÁLLÍTÁS. *Ha $k \geq 4$, akkor a (8)–(9) elem esetén a diszkrét gradiens operátor nulltere, az $N_{V_h(\Omega)}$ halmaz $(\sigma + 1)$-dimenziós, ahol σ a szinguláris pontok számát jelöli.*

A $k \geq 4$ feltétel lényeges; ha például $k = 2$ és $\Omega = [0,1]^2$, akkor standard triangularizáció esetén $\dim N_{V_h(\Omega)} = 6$ (azaz $\dim N_{V_h(\Omega)} = \sigma + 4$), míg a jól ismert „criss-cross" rács (ld. [3]) esetén $\dim N_{V_h(\Omega)} = \sigma + 1$ érvényes.

Az állítás szerint szinguláris pontok jelenléte esetén a diszkrét gradiens nulltere nem részhalmaza a folytonos gradiens operátor nullterének, a bevezetésben már említett jelenséggel találkozunk: a megoldandó lineáris egyenletrendszer nulltere többdimenziós, „energiamentes" nem konstans nyomásfüggvények vannak jelen.

[10]-ben a szerzők definiáltak egy függvényt, amely azt méri, hogy egy nem szinguláris x_0 belső rácspont mennyire közel van ahhoz, hogy szinguláris pont legyen. A 3.1. Definíció jelöléseivel az $R(x_0)$ függvényt a következő módon definiáljuk:

$$R(x_0) := \max\{|\Theta_i + \Theta_j - \pi|, \quad \text{ahol} \quad 1 \leq i, j \leq n,\ i - j = 1 \mod n\}.$$

Legyen $\{\mathcal{T}_h\}$, $0 < h \leq 1$ triangularizációk egy családja. A (8)–(9) elem $k \geq 4$ esetén csak akkor stabil, ha létezik egy olyan h-tól független δ konstans, hogy

$$\min\{R(x_0) : x_0 \in \Omega \backslash \Gamma \text{ nem szinguláris rácspont } \mathcal{T}_h\text{-ból}\} \geq \delta > 0$$

teljesül (ld. [10]).

Ha a diszkretizációs paraméter csökkenésével egy nem szinguláris pont tart a szinguláris helyzethez, akkor a stabilitás nem teljesül.

A criss-cross rács számos szinguláris pontot tartalmaz, de a standard rácsgeneráló programok által készített rácsokban is gyakran megfigyelhetők szinguláris, vagy közel-szinguláris pontok, így többen foglalkoztak azzal, hogyan lehetne ezt a rácsra vonatkozó kellemetlen feltételt kiküszöbölni. Egy lehetséges megoldás

a $V_h(\Omega)$ tér nemkonform bővítése: a sebesség folytonosságát a szomszédos háromszögek közös oldalain csak bizonyos pontokban követeljük meg. Ha ezeket a pontokat a k-adrendű esetben az adott oldalon definiált k-adfokú Legendre-polinom gyökeinek (a k-adrendű Gauss-Legendre-pontoknak) választjuk, akkor a (6) bilineáris funkcionállal normát definiálhatunk ezen a kibővített téren.

3.2. Definíció. A k-adrendű Gauss–Legendre-elem:

$$P_h(\Omega) = \left\{p_h \in L^2(\Omega),\ p_h|_\Delta \in \mathbb{P}_{k-1}(\Delta),\ \Delta \in \mathcal{T}_h\right\}, \tag{10}$$

$$V_h^{nc}(\Omega) = \left\{\vec{v}_h \in (L^2(\Omega))^2,\ \vec{v}_h|_\Delta \in (\mathbb{P}_k(\Delta))^2,\ \text{és } \vec{v}_h \text{ folytonos a } \Delta \right.$$
$$\left. \text{háromszög összes Gauss–Legendre-pontjában } \Delta \in \mathcal{T}_h\right\}, \tag{11}$$

$$\text{a norma } V_h\text{-n: } |\vec{v}_h|_{1,h,\Omega} := \left(\sum_{\Delta\in\mathcal{T}_h} |\vec{v}_h|^2_{1,\Delta}\right)^{1/2}.$$

Megjegyzés.

1. Ebben az esetben a homogén peremfeltétel helyett a sebesség v_ℓ, $\ell = 1, 2$, koordináta függvényeire

$$\int_{\Gamma_j} q v_\ell \mathrm{d}s = 0, \quad q \in \mathbb{P}_{k-1}(\Gamma_j)$$

teljesül $\forall \Gamma_j \subset \partial\Delta \cap \partial\Omega$, $\forall \Delta \in \mathcal{T}_h$ esetén.

2. $V_h^{nc}(\Omega) \not\subset (H^1(\Omega))^2$ (az elem nem konform), de a sebességek folytonosak a Gauss–Legendre-pontokban, ezért az $|.|_{1,h,\Omega}$ szeminorma normát definiál $V_h^{nc}(\Omega)$-n.
3. A (10)-(11) végeselem család az ismert Crouzeix–Raviart ($k = 1$), Fortin–Soulie ($k = 2$) és Crouzeix–Falk ($k = 3$) elemek általánosítása. A $k = 4, 6$ esetek vizsgálata [4]-ben szerepel.

Páratlan k esetén a $V_h^{nc}(\Omega)$ tér elemei egyértelműen leírhatóak, ha szabadsági fokoknak az alábbi csomópontokban felvett függvényértékeket választjuk: a háromszögek belsejében egyenletesen elosztunk $(k-2)(k-1)/2$ pontot, a maradék $3k$ pontot pedig a háromszög oldalain, a k-adfokú Legendre-polinom zérushelyeinél helyezzük el. Páros k esetén azonban létezik olyan k-adfokú polinom, amely a háromszög oldalain csak a k-adrendű Gauss–Legendre-pontokban tűnik el.

3.3. Definíció. Páros k esetén a k-adrendű nemkonform buborék függvény olyan (az adott háromszögön) definiált polinom, mely a háromszög minden oldalán a k-adrendű Legendre-polinommal egyenlő:

$$B_{n,\Delta}^{(k)} = \frac{1}{2}\left(\sum_{i=1}^{3} \mathsf{P}_k^{(0,0)}(1-2\lambda_i) - 1\right), \tag{12}$$

ahol $\mathsf{P}_k^{(0,0)}$ jelöli az 1 főegyütthatójú, a $[-1,1]$-en értelmezett k-adrendű Legendre-polinomot, λ_i, $i=1,2,3$, pedig a Δ-n definiált baricentrikus koordináták.

Megjegyzés.

1. $k \geq 4$ esetén a definícióban leírt tulajdonsággal nem csak a (12) alakú függvények rendelkeznek, hanem minden $B_{n,\Delta}^{(k)} + B_{c,\Delta}^{(k)}$ alakú függvény. Itt $B_{c,\Delta}^{(k)}$ egy k-adrendű konform buborék függvény: $B_{c,\Delta}^{(k)} = \lambda_1\lambda_2\lambda_3 q_{k-3}$, ahol q_{k-3} tetszőleges, a Δ háromszögön definiált $(k-3)$-adfokú polinom.
2. Páratlan k esetén a (12) függvény a háromszög oldalain azonosan nulla.
3. A $k=2$ esetben

$$B_{n,\Delta}^{(2)} = \frac{1}{2}\left\{\sum_{i=1}^{3} \mathsf{P}_2^{(0,0)}(1-2\lambda_i) - 1\right\} = 3\sum_{i=1}^{3}\lambda_i^2 - 2,$$

ami éppen a [6]-ban használt buborék függvény, míg $k=4$ és $k=6$ esetén $B_{n,\Delta}^{(k)}$ a [4]-ben használt buborék függvényektől csak egy konform tagban különbözik.

3.2. ÁLLÍTÁS. *Páros k esetén a (12) segítségével $V_h^{nc}(\Omega)$ a következő módon is leírható:*

$$V_h^{nc}(\Omega) = V_h(\Omega) + \left\{\vec{v},\ \vec{v}|_\Delta = \begin{pmatrix}\alpha_\Delta\\ \beta_\Delta\end{pmatrix} B_{n,\Delta}^{(k)},\ \alpha_\Delta,\beta_\Delta \in \mathbb{R},\ \Delta \in \mathcal{T}_h\right\}. \tag{13}$$

3.1. TÉTEL. *Ha k páros, akkor a (10)–(11) végeselem pár esetén a diszkrét gradiens operátor nulltere egydimenziós, azaz a*

$$b(\vec{v}_h, p_h) = 0 \qquad \forall \vec{v}_h \in V_h^{nc}(\Omega) \tag{14}$$

összefüggés csak konstans p_h esetén teljesül.

Bizonyítás. Legyen először $k=2$, és tegyük fel, hogy $p_h \in P_h(\Omega)$-ra (14) teljesül. Legyen $\Delta \in \mathcal{T}_h$ egy olyan háromszög, melyen p_h nem azonosan nulla. Ekkor $\vec{v}_h|_\Delta = (B_{n,\Delta}^{(2)}, 0)$, $\vec{v}_h|_{\Omega\setminus\Delta} \equiv 0$ választással, felhasználva, hogy $\frac{\partial p_h}{\partial x_1}$ konstans, (14)-ből következik, hogy

$$0 = \int_\Omega p_h \mathrm{div}\vec{v}_h \mathrm{d}x = -\int_\Delta \frac{\partial p_h}{\partial x_1} B_{n,\Delta}^{(2)} \mathrm{d}x = -\frac{\partial p_h}{\partial x_1}\int_\Delta B_{n,\Delta}^{(2)}\mathrm{d}x = \frac{\partial p_h}{\partial x_1}\cdot\frac{1}{4},$$

így $\frac{\partial p_h}{\partial x_1} \equiv 0$. Hasonlóan adódik, hogy $\frac{\partial p_h}{\partial x_2} \equiv 0$, így p_h konstans a Δ háromszög fölött. Annak igazolásához, hogy p_h konstans az egész tartományon, legyen Δ_1 és Δ_2 két közös oldallal rendelkező háromszög, és teljesüljön $p_h|_{\Delta_i} \equiv c_i \in \mathbb{R}$, $i=1,2$. Legyen v a skalár eset Lagrange-bázisának az az eleme, melynek értéke a

két háromszög közös oldalának felezőpontjában nem nulla. A $\vec{v}_{h,1}|_{\Delta_1\cup\Delta_2} = (v,0)$, $\vec{v}_{h,1}|_{\Omega\setminus(\Delta_1\cup\Delta_2)} \equiv 0$ és $\vec{v}_{h,2}|_{\Delta_1\cup\Delta_2} = (0,v)$, $\vec{v}_{h,2}|_{\Omega\setminus(\Delta_1\cup\Delta_2)} \equiv 0$ függvényekkel felírva a $b(\vec{v}_h, p_h) = 0$ egyenletet kapjuk, hogy $c_1 = c_2$.

$k \geq 4$ esetén a részletesebb bizonyítást lásd [1]-ben és [2]-ben. A bizonyítás vázlata: a 3.2. állítást felhasználva előbb leírjuk az $N_{V_h(\Omega)}$ teret, majd belátjuk, hogy minden $p_h \in N_{V_h(\Omega)}$-ra

$$b(\vec{v}_h, p_h) \neq 0$$

teljesül a $\vec{v}_h|_\Delta = (B^{(k)}_{n,\Delta}, 0)$, $\vec{v}_h|_{\Omega\setminus\Delta} \equiv 0$ vagy $\vec{v}_h|_\Delta = (0, B^{(k)}_{n,\Delta})$, $v_h|_{\Omega\setminus\Delta} \equiv 0$ függvények valamelyikével, ahol Δ egy tetszőleges háromszög p_h tartójából.

Az $N_{V_h(\Omega)}$ teret egy olyan bázisával írjuk le, amely a konstans függvény mellett σ darab olyan függvényt tartalmaz, amelyek mindegyike hozzárendelhető a rács egy szinguláris pontjához oly módon, hogy csak a szinguláris pontot tartalmazó háromszögeken vesz fel nullától különböző értékeket.

Vizsgáljuk a $b(\vec{v}_h, p_h) = 0$ egyenletet először olyan $\vec{v}_h$ függvényekre, melyek csak egy rögzített Δ háromszög belsejében vesznek fel nullától különböző értékeket. Ekkor $\vec{v}_h|_\Delta = \lambda_1\lambda_2\lambda_3\vec{q}_{k-3}$, ahol λ_i, $i = 1,2,3$, a Δ-beli baricentrikus koordináták és $\vec{q}_{k-3}$ a Δ háromszög fölött definiált tetszőleges $(k-3)$-adfokú polinom. A $b(\vec{v}_h, p_h) = 0$ egyenletből parciális integrálás után kapjuk, hogy $\frac{\partial p_h}{\partial x_1}$ és $\frac{\partial p_h}{\partial x_2}$ olyan $(k-2)$-edfokú polinomok, melyek a Δ háromszög fölött a $\lambda_1\lambda_2\lambda_3$ súlyfüggvényre nézve ortogonálisak $\mathbb{P}_{k-3}(\Delta)$-ra. Felhasználva a háromszögek fölött ortogonális polinomrendszer leírását [8] belátható, hogy adott szinguláris pont esetén a hozzátartozó bázis függvény minden olyan Δ háromszögben, melynek a szinguláris pont csúcspontja, a $\mathsf{P}^{(0,2)}_{k-1}(1-2\lambda_i)$ konstansszorosával egyenlő. Itt $\mathsf{P}^{(0,2)}_{k-1}$ a $[-1,1]$-en értelmezett $(0,2)$ paraméterű Jacobi-polinom, λ_i pedig az a baricentrikus koordináta Δ-n, amelynek az értéke a szinguláris pontban 1. Részletesebben, a különböző típusú perem szinguláris pontok és a belső szinguláris pont esetén:

A) Legyen az S_0 I-es típusú perem szinguláris pont. Ekkor a triangularizáció egyetlen háromszögének (legyen ez Δ_1) csúcspontja S_0 (ld. 1. ábra, itt az S_0S_1 és S_0S_2 szakaszok a tartomány peremén helyezkednek el). Az S_0 szinguáris ponthoz rendelt eleme a bázisnak:

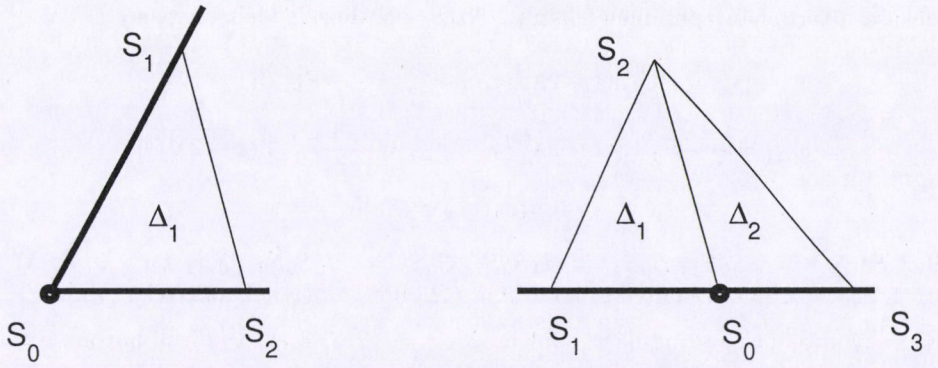
$$q_h|_{\Delta_1} = \mathsf{P}^{(0,2)}_{k-1}(1-2\lambda_3), \quad q_h|_{\Omega\setminus\Delta_1} \equiv 0,$$

ahol λ_3 az a baricentrikus koordináta Δ_1-ben, amelynek S_0-ban az értéke 1.

B) Legyen S_0 II. típusú perem szinguláris pont. Ekkor S_0 a triangularizáció két háromszögének (Δ_1 és Δ_2) csúcspontja, és a két háromszög S_0 pontnál lévő szögének összege π (ld. 1. ábra, itt az S_1S_3 szakasz része a tartomány peremének). Az S_0 ponthoz tartozó eleme a bázisnak

$$q_h|_{\Delta_1} = \mathsf{P}^{(0,2)}_{k-1}(1-2\lambda^{(1)}_3), \qquad q_h|_{\Delta_2} = -\frac{1}{t_0}\mathsf{P}^{(0,2)}_{k-1}(1-2\lambda^{(2)}_3), \qquad q_h|_{\Omega\setminus(\Delta_1\cup\Delta_2)} \equiv 0,$$

ahol $\vec{S_0S_3} = -t_0\vec{S_0S_1}$, $t_0 > 0$, és $\lambda^{(1)}_3$, ill. $\lambda^{(2)}_3$ az a baricentrikus koordináta Δ_1-ben, ill. Δ_2-ben, amelynek az értéke S_0-ban 1.



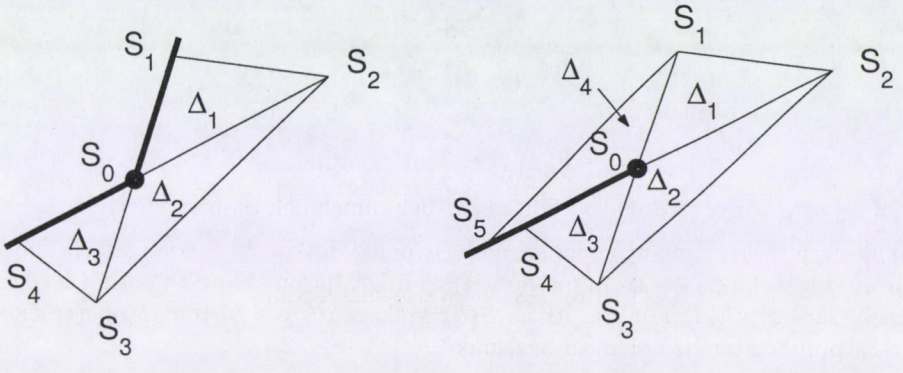
1. ábra. I-es és II-es típusú perem szinguláris pont

C) Legyen S_0 III. típusú perem szinguláris pont. Ekkor S_0 a triangularizáció 3 háromszögének $(\Delta_1, \Delta_2, \Delta_3)$ csúcspontja (ld. 2. ábra, ahol az S_0S_1 és S_0S_4 szakaszok a tartomány peremén vannak). Az S_0 ponthoz rendelt függvény:

$$q_h|_{\Delta_1} = \mathrm{P}_{k-1}^{(0,2)}(1-2\lambda_3^{(1)}), \qquad q_h|_{\Delta_2} = -\frac{1}{t_0}\mathrm{P}_{k-1}^{(0,2)}(1-2\lambda_3^{(2)}),$$

$$q_h|_{\Delta_3} = \frac{1}{t_0 t_1}\mathrm{P}_{k-1}^{(0,2)}(1-2\lambda_3^{(3)}), \qquad q_h|_{\Omega\setminus(\Delta_1\cup\Delta_2\cup\Delta_3)} \equiv 0,$$

ahol $\vec{S_0S_3} = -t_0\vec{S_0S_1}$, $\vec{S_0S_4} = -t_1\vec{S_0S_2}$, $t_0, t_1 > 0$, és $\lambda_3^{(i)}$, $i = 1,2,3$, az a baricentrikus koordináta Δ_i-ben, amelynek az értéke S_0-ban 1.



2. ábra. III. és IV. típusú perem szinguláris pont

D) Legyen S_0 IV. típusú perem szinguláris pont. Ekkor az S_0 pont a triangularizáció 4 háromszögének $(\Delta_i,\ i = 1,2,3,4)$ csúcspontja (ld. 2. ábra, ahol az $S_0S_4S_5$

szakasz a tartomány peremén fekszik). Az S_0 ponthoz rendelt függvény:

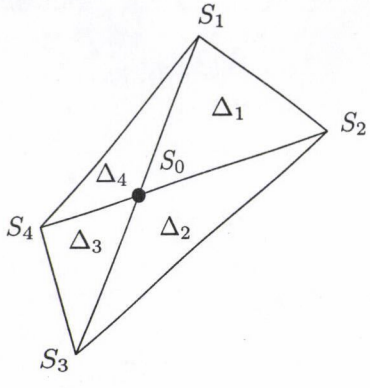
$$q|_{\Delta_1} = \mathsf{P}_{k-1}^{(0,2)}(1-2\lambda_3^{(1)}), \quad q|_{\Delta_2} = -\frac{1}{t_0}\mathsf{P}_{k-1}^{(0,2)}(1-2\lambda_3^{(2)}),$$
$$q|_{\Delta_3} = \frac{1}{t_0 t_1}\mathsf{P}_{k-1}^{(0,2)}(1-2\lambda_3^{(3)}), \quad q|_{\Delta_4} = -\frac{1}{t_2}\mathsf{P}_{k-1}^{(0,2)}(1-2\lambda_3^{(4)}),$$
$$q|_{\Omega\setminus(\Delta_1\cup\Delta_2\cup\Delta_3}\equiv 0,$$

ahol $S_0\vec{S}_3 = -t_0 S_0\vec{S}_1$, $S_0\vec{S}_4 = -t_1 S_0\vec{S}_2$, $S_0\vec{S}_5 = -t_2 S_0\vec{S}_2$, $t_0, t_1, t_2 > 0$, és $\lambda_3^{(i)}$, $i = 1,2,3,4$, az a baricentrikus koordináta Δ_i-ben, amelynek az értéke S_0-ban 1.

E) Legyen S_0 belső szinguláris pont és Δ_i, $i = 1,2,3,4$, az S_0 körüli háromszögek (lásd a 3. ábrát). Ekkor az S_0 ponthoz rendelt függvény:

$$q|_{\Delta_1} = \mathsf{P}_{k-1}^{(0,2)}(1-2\lambda_3^{(1)}), \quad q|_{\Delta_2} = -\frac{1}{t_0}\mathsf{P}_{k-1}^{(0,2)}(1-2\lambda_3^{(2)}),$$
$$q|_{\Delta_3} = \frac{1}{t_0 t_1}\mathsf{P}_{k-1}^{(0,2)}(1-2\lambda_3^{(3)}), \quad q|_{\Delta_4} = -\frac{1}{t_1}\mathsf{P}_{k-1}^{(0,2)}(1-2\lambda_3^{(4)}),$$
$$q|_{\Omega\setminus(\Delta_1\cup\Delta_2\cup\Delta_3\cup\Delta_4)}\equiv 0,$$

ahol $S_0\vec{S}_3 = -t_0 S_0\vec{S}_1$, $S_0\vec{S}_4 = -t_1 S_0\vec{S}_2$, $t_0, t_1 > 0$, és $\lambda_3^{(i)}$, $i = 1,\ldots,4$, azok a baricentrikus koordináták Δ_i-ben, $i = 1,2,3,4$, melyek értéke az S_0-ban 1.



3. ábra. Egy belső szinguláris pont. □

4. A páros rendű Gauss–Legendre-elemek stabilitása

Páros $k \geq 2$ értékek esetén a stabilitás a [11]-ben leírt makroelem módszer nemkonform esetre készített módosításával bizonyítható (részletesen lásd [2]-ben).

Itt alapvető szerepet játszik az a tény, hogy a páros rendű Gauss–Legendre-elemek esetén a diszkrét gradiens nulltere – a rács esetleges szingularitásától függetlenül – egydimenziós. Először (ugyanúgy, mint [11]-ben) definiáljuk a makroelemeket, ill. a makroelemek ekvivalenciáját.

4.1. Definíció. Egy makroelem $\mathcal{T}_h$-beli szomszédos háromszögek uniója. Az M makroelem ekvivalens az $\hat{M}$ referencia makroelemmel, ha létezik olyan $F_M : \hat{M} \to M$ leképezés, melyre az alábbi feltételek teljesülnek:

1. F_M folytonos és kölcsönösen egyértelmű,
2. $F_M(\hat{M}) = M$,
3. ha $\hat{M} = \bigcup_{j=1}^{m} \hat{\Delta}_j$, ahol $\hat{\Delta}_j$, $j = 1, \ldots, m$, az $\hat{M}$-et alkotó háromszögek, akkor az M makroelemet a $\Delta_j = F_M(\hat{\Delta}_j)$, $j = 1, \ldots, m$, háromszögek alkotják,
4. $F_{M|\Delta_j} = F_{\Delta_j} \circ F_{\hat{\Delta}_j}^{-1}$, $j = 1, \ldots, m$, ahol $F_{\hat{\Delta}_j}$ és F_{Δ_j} a referencia háromszöget $\hat{\Delta}_j$-re, ill. Δ_j-re leképező affin transzformációk.

A stabilitás igazolásához makroelemek olyan $\mathcal{E}_{\hat{M}_i}, i = 1, \ldots, n$, $n \geq 1$, ekvivalencia osztályait kell definiálnunk, amelyekre a következő két feltétel teljesül:

1. tetszőleges h esetén a $\mathcal{T}_h$-beli háromszögek összecsoportosíthatóak makroelemekké úgy, hogy az így kapott $\mathcal{M}_h$ makroelem-felosztás minden $M \in \mathcal{M}_h$ eleme besorolható valamelyik $\mathcal{E}_{\hat{M}_i}$, $i = 1, \ldots, n$ makroelem-osztályba,
2. minden $M \in \mathcal{E}_{\hat{M}_i}$, $i = 1, \ldots, n$ esetén az

$$N_M^{nc} := \{p_h \in P_h(M) : b(\vec{v}_h, p_h) = 0 \quad \forall \vec{v}_h \in V_h^{nc}(M)\}$$

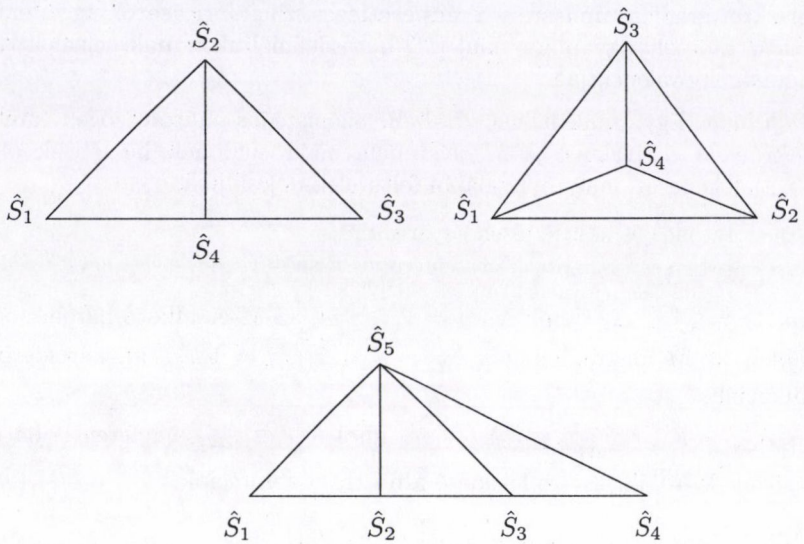
tér egydimenziós.

4.1. TÉTEL. *Ha a fenti két feltétel teljesül, akkor a* (10)–(11) *elem inf-sup stabil.*

A tétel bizonyítása [2]-ben található.

4.2. TÉTEL. *Páros $k \geq 2$ esetén a* (10)–(11) *elem inf-sup stabil.*

Bizonyítás. Esetünkben 3 makroelem osztályt definiálunk; $\mathcal{E}_{\hat{M}_1}$-be tartoznak azok a makroelemek, amelyek két szomszédos (közös oldallal rendelkező) háromszögből állnak, $\mathcal{E}_{\hat{M}_2}$-be azok a makroelemek, amelyeket 3 olyan háromszög alkot, amelyek közül bármely kettőnek van közös oldala, $\mathcal{E}_{\hat{M}_3}$-at a 3 szomszédos háromszögből álló, nem az $\mathcal{E}_{\hat{M}_2}$-be tartozó makroelemek alkotják. Az $\hat{M}_1$, $\hat{M}_2$, $\hat{M}_3$ referenciaelemek (melyekre a megfelelő osztályok elemei folytonos, kölcsönösen egyértelmű módon leképezhetőek):



A $\mathcal{E}_{\hat{M}_i}$, $i = 1, 2, 3$ osztályok teljesítik az 1. makroelem feltételt, a 2. feltételben leírt állítás pedig következik 3.1. Tételből. Mivel a két makroelem feltétel teljesül, a (10)–(11) végeselem stabil. □

Megjegyzés.

1. Itt fontos szerepe van a diszkrét gradiens nulltere dimenziójának. Míg a konform Scott–Vogelius-elemek esetén ha a triangularizációban egy közel szinguláris pont tart a szinguláris helyzethez, a megfelelő nullterek nem folytonos módon változnak (a határhelyzetben a nulltér dimenziója eggyel nagyobb), addig a Gauss–Legendre-elemek esetén a nulltér mindig egydimenziós, csak a konstansfüggvényt tartalmazza.
2. Páratlan rendű Gauss–Legendre-elemek esetén a stabilitás ugyanezen makroelem osztályok választásával nem igazolható. Be lehet látni, hogy a páratlan $k \geq 3$ értékekre az $\mathcal{E}_{\hat{M}_1}$ osztályba tartozó M makroelemek esetén az N_M^{nc} térnek van legalább egy nem konstans eleme (ld. [1]). A $k = 3$ esetet [5]-ben vizsgálták, ott bizonyos triangularizációkra megmutatták az elem stabilitását és sejtésként megemlítik, hogy az elem tetszőleges triangularizáció esetén stabil.

Hivatkozások

[1] Á. BARAN: *A high-order non-conforming finite element family*, PhD értekezés, Debreceni Egyetem, Informatikai Kar, 2007.

[2] Á. BARAN, G. STOYAN: *Gauss-Legendre-elements: a stable higher order non-conforming finite element family*, Computing **79**, no. 1, 1–21 (2007).

[3] F. BREZZI, M. FORTIN: *Mixed and Hybrid Finite Element Methods*, Springer–Verlag New York, 1991.

[4] Y. CHA, M. LEE, S. LEE: *Stable nonconforming methods for the Stokes problem*, Applied Mathematics and Computation **114**, 155–174 (2000).

[5] M. CROUZEIX, R. S. FALK: *Nonconforming finite elements for the Stokes problem*, Mathematics of Computation **186**, 437–456 (1989).

[6] M. FORTIN, M. SOULIE: *A non-conforming piecewise quadratic finite element on triangles*, Int. J. Numer. Methods Eng. **19**, 505–520 (1983).

[7] V. John, G. Matthies, Higher order finite element discretizations in a benchmark problem for incompressible flows, *International Journal for Numerical Methods in Fluids* **37**, 885–903, (2001).

[8] T. Koornwinder, Two-variable analogues of the classical orthogonal polynomials. In: Theory and Application of Special Functions (R. Askey ed.), pp 435–495, Academic Press, 1975.

[9] G. MATTHIES, L. TOBISKA: *Inf-sup stable non-conforming finite elements of arbitrary order on triangles*, Numerische Mathematik **102**, 293–309 (2005).

[10] L.-R. SCOTT, M. VOGELIUS: *Norm estimates for a maximal right inverse of the divergence operator in spaces of piecewise polynomials*, Modélisation Mathématique et Analyse Numérique **19**, 111–143 (1985).

[11] R. STENBERG: *Analysis of mixed finite element methods for the Stokes problem: a unified approach*, *Math. of Comp.* **165**, 9–23 (1984).

[12] G. STOYAN, Á. BARAN: *Crouzeix-Velte decompositions for higher-order finite elements*, *Comp. Math. with Appls.* **51**, 967–986 (2006).

[13] M. SCHÄFER, S. TUREK: *The benchmark problem "Flow around a cylinder"*. In: E. H. Hirschel, editor, Flow Simulation with High-Performance Computers II vol. **52** of *Notes on Numerical Fluid Mechanics*, 547–566, (1996).

BARAN ÁGNES
Debreceni Egyetem, Informatikai Kar
Alkalmazott Matematika és Valószínűségszámítás Tanszék
4010 Debrecen, Pf. 12.
baran.agnes@inf.unideb.hu

A HIGH-ORDER NON-CONFORMING FINITE ELEMENT FAMILY FOR THE SOLUTION OF THE TWO-DIMENSIONAL STOKES PROBLEM

ÁGNES BARAN

In this paper we describe a triangular non-conforming finite element family for the two-dimensional Stokes problem. Similarly to the conforming element pair defined by Scott and Vogelius, pressure and velocity are approximated trianglewise by polynomials of order $k-1$ and k, respectively. The continuity of the discrete velocity on the common sides of the triangles, unlike the Scott-Vogelius element, is required at particular points only. The finite element pair is defined for all $k \geq 1$ and it is a generalization of low order ($k = 1, 2, 3$) cases. We show that for even k the finite element pair can be obtained from the Scott-Vogelius element by adding trianglewise a non-conforming bubble function to the local basis of the velocity space. The bubble function removes the algebraic dicontinuity of the Scott-Vogelius elements, i.e. the presence of the "energy-free" discrete pressure. We show that the element pair is stable for even k.

Alkalmazott Matematikai Lapok **26** (2009), 207-222.

HÁLÓNÉLKÜLI MÓDSZEREK ÉS ALKALMAZÁSUK A STOKES-PROBLÉMÁRA

GÁSPÁR CSABA

A cikkben a kétdimenziós időfüggetlen Stokes-egyenletrendszer egy lehetséges numerikus megoldását vázoljuk. A módszer az ismert Uzawa-algoritmuson alapuló nyomáskorrekciós módszer hálómentes (rácsmentes) változata: a nyomáskorrekciós módszerben fellépő Poisson-egyenleteket hálómentes módszerrel oldjuk meg. Így az áramlási tartományt sem véges elemekkel, sem végesdifferenciás rácshálózattal diszkretizálni nem szükséges. Az alkalmazott sémák a radiális bázisfüggvények módszerének lokális változatából adódnak. A kapott módszer számításigénye viszonylag csekély, és a módszer könnyen multigrid környezetbe ágyazható. A módszert egy egyszerű numerikus tesztfeladaton keresztül szemléltetjük.

1. A Stokes-egyenletek

Az összenyomhatatlan folyadék mozgását matematikailag a következő differenciálegyenlet-rendszer írja le (*Navier–Stokes-egyenletek*):

$$\begin{aligned} \operatorname{div} \mathbf{u} &= 0 \qquad \text{(folytonossági egyenlet)}, \\ \frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \frac{1}{\rho}\operatorname{grad} p - \nu\Delta\mathbf{u} &= \mathbf{a} \qquad \text{(momentumegyenletek)}, \end{aligned} \tag{1}$$

ahol $\mathbf{u} = (u, v, w)$ a sebesség (vektor), p a nyomás (skalár), ρ a folyadék sűrűsége, melyről feltesszük, hogy térben és időben változatlan, ν pedig a kinematikai viszkozitás, melyet szintén konstansnak tekintünk. Más szóval, a sűrűség és viszkozitás helyi és időbeli változásaitól eltekintünk. Az (1) egyenletben ∇ a nabla operátort jelöli $\left(\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right)\right)$, Δ pedig a Laplace-operátort $\left(\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right)$. A momentumegyenletek jobb oldalán álló $\mathbf{a}$ vektorfüggvény a külső erők által létrehozott gyorsulást írja le.

Ha időben változatlan (permanens) folyamatokkal foglalkozunk, akkor a fenti egyenletekben az idő szerinti deriváltak mind 0-val egyenlők.

Az (1) egyenlet speciális esete, amikor az $(\mathbf{u}\cdot\nabla)\mathbf{u}$ konvektív deriváltakat elhanyagoljuk. Ez *lassú áramlások* és/vagy *nagy viszkozitás* mellett tehető meg.

Az áramlást ekkor a *Stokes-egyenletek* írják le:

$$\operatorname{div} \mathbf{u} = \mathbf{a}, \tag{2}$$
$$\frac{\partial \mathbf{u}}{\partial t} - \nu \Delta \mathbf{u} + \frac{1}{\rho} \operatorname{grad} p = 0.$$

A (2) egyenletek *lineáris* differenciálegyenletek, míg az (1) egyenletek nemlineárisak. Mind elméleti, mind numerikus szempontból a (2) egyenletek linearitása igen nagy előnyt jelent: ezenfelül az (1) egyenletre léteznek olyan iterációs módszerek, melyek minden lépésében egy-egy (2) típusú lineáris egyenletet kell megoldani.

1.1. A peremfeltételek problémája

Más peremfeltételt kell megadni a sebességkomponensek, és mást a nyomás esetén; továbbá a perem egyes részein is változik a peremfeltétel jellege. A peremet bontsuk három különböző, nem feltétlen összefüggő részre: beáramlási perem, kiáramlási perem, szilárd perem (fal). Az egyes részeken a következő adatokat szokás peremfeltételként megadni:

Beáramlási perem: Itt mindhárom sebességkomponenst (u, v, w) meg kell adni, míg a p nyomásra nézve semmilyen előírást nem teszünk. A gyakorlatban általában a belépő *vízhozamot* ismerjük (egységnyi idő alatt beáramló vízmennyiség), a sebességkomponensek eloszlását a beáramlási perem mentén nem. Külön feladat tehát a sebességeloszlás előállítása az adott hozam mellett. Legegyszerűbb esetben konstans nagyságú, a beáramlási felületre merőleges (normális) irányú sebességet írunk elő: ez a valóságot csak durván írja le, mivel nem veszi figyelembe a falsúrlódás okozta sebességcsökkenést. Finomabb modellezés tehető adott (pl. kvadratikus) sebességprofil meghatározásával. Ez azonban bonyolultabb alakú felületnél nem túl egyszerű, de jelentős hatása csak a beáramlási perem közvetlen környezetében van.

Szilárd perem (fal): Itt a normális irányú sebességkomponens mindenképp zérus (mivel a falon keresztül nincs áramlás), a tangenciális komponensre pedig az alábbi előírások valamelyikét tesszük:

- *Csúszó perem*: a tangenciális irányú sebességkomponensek normális irányú deriváltja zérus.
- *Tapadó perem*: a tangenciális irányú sebességkomponensek zérussal egyenlők.
- *Félig tapadó perem*: az előbbi kettő közti átmenet: matematikailag ez a tangenciális sebességkomponensekre nézve egy harmadfajú peremfeltételt jelent.

Technikailag a legegyszerűbben a tapadó perem modellezhető. A nyomásra peremfeltételt itt sem adunk.

Kiáramlási perem: Itt a nyomásértéket szokás megadni. A sebességkomponenseket nem írjuk elő, de feltesszük, hogy a sebességkomponensek normális (kilépő) irányban nem változnak, tehát normális irányú deriváltjuk zérus. Egy másik lehetőség, hogy a kiáramlási perem mentén (éppúgy, mint a beáramlási perem mentén) a nyomásra nem adunk peremfeltételt, a sebességkomponenseket pedig explicite előírjuk. Ha azonban a kiáramlási perem nem összefüggő, akkor sokszor nem ismerjük a kilépő sebességeket még közelítően sem: a feladat részben épp az lehet, hogy a kilépő hozam milyen arányban oszlik meg az egyes kiáramlási peremdarabok között.

A későbbiekben a *kétdimenziós*, stacionárius, külső erőhatásoktól mentes Stokes-féle áramlásokkal foglalkozunk. A megfelelő egyenletek alakja a szokásos primitív változókkal így a következő:

$$\begin{aligned} \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} &= 0, \\ \nu \Delta u &= \frac{1}{\rho} \frac{\partial p}{\partial x}, \\ \nu \Delta v &= \frac{1}{\rho} \frac{\partial p}{\partial y}. \end{aligned} \tag{3}$$

Az Ω áramlási tartomány teljes Γ peremén az u, v sebességkomponenseket írjuk elő. A p nyomásra nézve azt a globális feltételt írjuk elő, hogy a teljes áramlási tartományon vett integrálja zérussal legyen egyenlő. Ismeretes, hogy ha Ω korlátos és szakaszonként sima, akkor ennek a problémának a $H^1(\Omega) \times H^1(\Omega) \times L_2(\Omega)$ függvénytérben létezik éspedig egyetlen (általánosított) megoldása, feltéve, hogy az u, v-re tett peremfeltételek a $H^{1/2}(\Gamma)$ függvénytérbe tartoznak, és kielégítik a folytonossági egyenletből következő $\int_\Gamma \mathbf{u} \cdot \mathbf{n}\, d\Gamma = 0$ feltételt. A megoldás egyúttal minimalizálja a

$$\int_\Omega \left(\|\mathrm{grad}\ u\|^2 + \|\mathrm{grad}\ v\|^2 \right)\, d\Omega$$

kvadratikus funkcionált az adott peremfeltétel és a $\mathrm{div}\,\mathbf{u} = 0$ feltétel mellett: a Lagrange-multiplikátor épp a p nyomás. A továbbiakban a megoldhatósági kérdésekkel nem foglalkozunk.

2. A nyomáskorrekciós módszer

A (3) Stokes-egyenletek analitikus megoldása általában – egészen ritka speciális esetektől eltekintve – reménytelen, így valamilyen numerikus módszert kell alkalmazni. A Stokes-egyenletekre több jól működő numerikus megoldási módszer ismeretes. Itt a klasszikus *Uzawa-algoritmust* alkalmazzuk, mely az (u, v, p)

függvényhármast az alábbi rekurzióval közelíti:

$$\begin{aligned} \Delta u_{k+1} &= \frac{1}{\nu\rho} \cdot \frac{\partial p_k}{\partial x}, \\ \Delta v_{k+1} &= \frac{1}{\nu\rho} \cdot \frac{\partial p_k}{\partial y}, \\ p_{k+1} &:= p_k - \omega\nu\rho \cdot \left(\frac{\partial u_{k+1}}{\partial x} + \frac{\partial v_{k+1}}{\partial y} \right), \end{aligned} \tag{4}$$

ahol az első két Poisson-egyenlethez Dirichlet-peremfeltételt csatolunk az adott peremértékekkel. Ismeretes, hogy minden, elég kis $\omega > 0$ iterációs paraméter mellett az iteráció a (3) Stokes-egyenlet (az adott peremfeltétel mellett érvényes) megoldásához konvergál.

Az algoritmus lényege tehát, hogy a Stokes-egyenletek megoldását Poisson-egyenletek sorozatának megoldására vezeti vissza.

Az Uzawa-algoritmus gyakorlati alkalmazásakor (4) első két Poisson-egyenletét nem szükséges pontosan megoldani, elég néhány (alul)relaxálási lépést alkalmazni (valamilyen diszkretizálási technika után). Így kapjuk az *egyszerű nyomáskorrekciós módszert*:

- Alkalmazzunk néhány (alul)relaxálást a diszkretizált momentumegyenletekre:
$$\Delta u = \frac{1}{\nu\rho} \cdot \frac{\partial p}{\partial x}, \qquad \Delta v = \frac{1}{\nu\rho} \cdot \frac{\partial p}{\partial y}.$$
- Korrigáljuk a nyomást a számított divergenciával:
$$p := p - \omega\nu\rho \cdot \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right),$$
ahol $\omega > 0$ egy elegendően kicsi iterációs paraméter, és ismételjük az eljárást az előző ponttól.

Kihasználhatjuk, hogy a pontos p nyomásfüggvény (a (3) egyenletek egyszerű következményeképp) az áramlási tartományban maga is kielégíti a Laplace-egyenletet (pontos peremfeltételek megadása azonban általában nem lehetséges). Így nyerjük a *simított nyomáskorrekciós módszert*. Itt az egyszerű nyomáskorrekció két lépése kiegészül még egy lépéssel:

- Alkalmazzunk néhány (alul)relaxálást a $\Delta p = 0$ egyenlet diszkrét megfelelőjére (az előző lépésben kapott nyomásból kiindulva), és módosítsuk a nyomást egy konstans hozzáadásával úgy, hogy a nyomás Ω-n vett integrálja 0 legyen:
$$p := p - \frac{1}{|\Omega|} \int_\Omega p \, d\Omega$$
(ahol $|\Omega|$ jelöli az tartomány területét).

Az iteráció minden lépésében tehát Poisson-egyenleteket szükséges közelítően megoldani. Numerikus szempontból fontos, hogy a Poisson-egyenletmegoldó algoritmus, ill. az azt közelítő relaxáció műveletigénye minél kevesebb legyen. Mindenekelőtt azonban a szóban forgó Poisson-egyenleteket diszkretizálni szükséges. Ennek szokásos és általánosan használt módja valamilyen (derékszögű vagy görbevonalú) rácsháló, és azon véges differenciasémák konstruálása; vagy pedig az áramlási tartomány végeselemes felbontása, és végeselem-technika alkalmazása. Az előbbi megközelítésben elterjedt az eltolt hálós (*staggered grid*) diszkretizáció [14], amikor a sebességkomponenseket és a nyomásokat más-más pontokban diszkretizáljuk (éspedig a sebességkomponenseket az egyes cellák élközéppontjaihoz, míg a nyomást a cellaközéppontokhoz rendeljük). A második megközelítésben ügyelni kell arra, hogy a sebességeket és a nyomást más-más módon kell közelíteni: csak ún. *stabil térpárok* jöhetnek számításba, melyek kielégítik a diszkrét inf-sup-feltételt [1], [15]. Mindkét megközelítés közös problémája azonban, hogy még a diszkrét egyenletek konstrukciója előtt derékszögű vagy görbevonalú rácsot, ill. végeselemhálót kell generálni. Nagyméretű és/vagy háromdimenziós probléma esetén ezek a feladatok az eredeti problémával összevethető bonyolultságúak, és speciális szoftverek nélkül alig megoldhatók. Jelen munkában a közelmúltban dinamikus fejlődésnek indult *hálónélküli* (meshfree, meshless) megközelítést alkalmazzuk, mely sem rács, sem háló generálását nem igényli, csak egy (a tartományban és annak peremén) kellően sűrű ponthalmaz generálását követeli meg: e ponthalmaznak azonban semmiféle struktúrával nem kell rendelkeznie.

3. Poisson-egyenletek hálónélküli megoldása

A hálónélküli módszerek konstrukciójának közös alapja valamilyen *szórt pontú interpoláció* alkalmazása. Először tehát röviden összefoglaljuk az idevonatkozó legfőbb fogalmakat. Végig kétdimenziós problémákat vizsgálunk azzal a megjegyzéssel, hogy az itt bemutatott módszerek értelemszerűen általánosíthatók három- vagy akár magasabb dimenziós problémákra is.

3.1. Szórt alappontú interpoláció

Legyen $\Omega_0 \subset \mathbf{R}^2$ korlátos tartomány, legyenek $x_1, x_2, \ldots, x_N \in \Omega_0$ adott, páronként különböző pontok, melyeken semmiféle (rács- vagy háló-) struktúrát nem tételezünk fel. Legyenek végül $f_1, f_2, \ldots, f_N \in \mathbf{R}$ adott számok. Keressünk olyan, minél simább $f : \Omega_0 \to \mathbf{R}$ függvényt, mely kielégíti az

$$f(x_k) = f_k \qquad (k = 1, 2, \ldots, N) \tag{5}$$

interpolációs feltételeket.

Nyilvánvaló, hogy a probléma ebben a megfogalmazásban alulhatározott, úgyhogy az interpolációs f függvényre további feltételeket kell tenni.

3.1. Példa. Definiáljuk az f interpolációs függvényt az alábbi súlyozott átlaggal (*inverz távolságok módszere* vagy *Shepard-módszer*):

$$f(x) := \frac{\sum_{j=1}^{N} f_j \cdot w_j(x)}{\sum_{j=1}^{N} w_j(x)},$$

ahol $x \in \mathbf{R}^2$ és $w_j(x) := \frac{1}{\|x-x_j\|^2}$ (itt $\|.\|$ jelöli az euklideszi normát). Ez a formula akkor értelmes, ha $x \neq x_k$ semelyik $k = 1, 2, \ldots, N$ indexre: ehelyett legyen $f(x_k) := f_k$ az alappontokban. Ismert, hogy f folytonos $\mathbf{R}^2$-n (sőt, folytonosan differenciálható), de az alappontokban mindkét parciális derivált eltűnik, ami az interpoláció pontosságát rontja.

A módszer realizálása (ellentétben a következőként mutatott módszerekkel) egyenletmegoldást nem igényel, műveletigénye minden egyes kiértékelés esetén $O(N)$. A módszer numerikusan stabil, az alappontoktól nagy távolságra az interpolált (helyesebben: extrapolált) érték közelítően az alapponti értékek egyszerű számtani átlaga. Jól alkalmazható pl. elliptikus egyenletek helyfüggő paramétereinek közelítésére (pl. hővezetési tényező, szivárgási tényező, vezetőképesség, mederfenékszintek stb.). Csekély pontossága miatt azonban hálómentes módszerek konstrukcióiban nem alkalmazzák.

Jelenleg a többdimenziós interpolációs problémák megoldásának valószínűleg a legnépszerűbb módszercsaládja a *radiális bázisfüggvények módszere* (RBF-módszer, [6], [7]). Ebben a megközelítésben, ha adottak az $f_1, f_2, \ldots, f_N \in \mathbf{R}$ függvényértékek az $x_1, x_2, \ldots, x_N \in \Omega_0$ interpolációs alappontokban, akkor az f interpolációs függvényt az alábbi alakban keressük:

$$f(x) := \sum_{j=1}^{N} \alpha_j \Phi_j(x - x_j), \tag{6}$$

ahol $\Phi_1, \Phi_2, \ldots, \Phi_N$ adott radiális (körszimmetrikus) bázisfüggvények, azaz a $\Phi_j(x)$ függvényértékek csak az $\|x\|$ (euklideszi) normától függnek. Az ismeretlen $\alpha_1, \alpha_2, \ldots, \alpha_N$ együtthatók pedig az *interpolációs egyenletekből* határozhatók meg:

$$\sum_{j=1}^{N} \alpha_j \Phi_j(x_k - x_j) = f_k \qquad (k = 1, 2, \ldots N) \tag{7}$$

feltéve, hogy ennek az egyenletrendszernek egyáltalán van megoldása (ez nem mindig teljesül [12]).

A $\Phi_j(x)$ radiális bázisfüggvények alkalmas megválasztásával sokféle konkrét módszerhez jutunk. A leggyakrabban használatos módszerek a következők (a radiális bázisfüggvényeket polárkoordinátákban megadva):

- *Multikvadrikus módszer* (method of multiquadrics, MQ-módszer [8]):

$$\Phi_j(r) := \sqrt{r^2 + c_j^2},$$

ahol a $c_1, c_2, \ldots c_N$ számok alkalmasan választott skálázó paraméterek. Egy lehetséges megválasztásuk:

$$c_k := \min_{j \neq k} \|x_k - x_j\|.$$

- *Vékony lemez módszer* (thin plate splines, TPS-módszer [2]):

$$\Phi_j(r) := r^2 \cdot \log r$$

minden $j = 1, 2, \ldots, N$ indexre.

- *Gauss-függvények*:

$$\Phi_j(r) := e^{-c_j^2 r^2}$$

($c_1, c_2, \ldots c_N$ itt is skálázó konstansok).

- Néha használatos még az alábbi radiális bázisfüggvény is (a perem-integrálegyenlet módszer duális reciprocitási módszerének nevezett technikában alkalmazták előszeretettel [13]):

$$\Phi_j(r) := 1 + |r|$$

minden $j = 1, 2, \ldots, N$ indexre.

Néha a (6) formula jobb oldalát kiegészítjük még néhány további taggal, tipikusan alacsony fokszámú polinomokkal:

$$f(x) := \sum_{j=1}^{N} \alpha_j \Phi_j(x - x_j) + \sum_{j=1}^{M} a_j p_j(x)$$

Ez esetben a (7) rendszer további egyenletekkel egészül ki (ortogonalitási feltételek):

$$\sum_{j=1}^{N} \alpha_j \Phi_j(x_k - x_j) + \sum_{j=1}^{M} a_j p_j(x_k) = f_k \qquad (k = 1, 2, ..., N),$$

$$\sum_{j=1}^{N} \alpha_j p_k(x_j) = 0 \qquad (k = 1, 2, ..., M).$$

A fenti radiális bázisfüggvények közös jellemzője, hogy bár a (7) interpolációs egyenletrendszer általában megoldható, de a rendszer numerikus szempontból nagyon kedvezőtlen. Miután e bázisfüggvények tartója nem korlátos, a rendszer mátrixa teljesen kitöltött mátrix, sokszor nemszimmetrikus és általában rosszul kondicionált: a kondíciószám tetszőlegesen nagy lehet, ha az interpolációs alappontok közt vannak egymáshoz közel eső pontok is. Ezért (7) megoldására általában – jobb híján – Gauss-eliminációt használnak: ennek műveletigénye $O(N^3)$, ami

megengedhetetlenül naggyá válik, ha az alappontok N száma nagy (a jelenlegi gyakorlatban: N meghaladja az ezres nagyságrendet). Ez a fenti módszerek alapvető hátránya, jóllehet, a módszerek interpolációs tulajdonságai nagyon jók: megmutatható, hogy az MQ-módszer (elég sima függvények esetén) *exponenciálisan* konvergál [11]. Hasonló eredmény érvényes a Gauss-függvényeken alapuló interpolációra is (míg a TPS-módszer konvergenciasebessége ennél alacsonyabb).

3.2. A radiális bázisfüggvények módszerére épülő hálónélküli módszerek

Modellfeladatként tekintsük a kétdimenziós Poisson-egyenletet Dirichlet-peremfeltétellel ellátva:

$$\Delta u = f \quad \Omega\text{-ban} \qquad u|_{\partial\Omega} = u_0. \tag{8}$$

Legyenek az $x_1, x_2, \ldots, x_M$ pontok az Ω tartományon, az $x_{M+1}, x_{M+2}, \ldots, x_{M+N}$ pontok pedig a $\partial\Omega$ peremen elhelyezve.

Aszerint, hogy magát az u megoldást vagy az f jobb oldalt approximáljuk RBF-módszerrel, a (8)-ra alkalmazott rácsnélküli RBF-módszerek két csoportra oszthatók:

1. *Kansa módszere* [9], [10]: Ekkor közvetlenül az u megoldást approximáljuk:

$$u(x) := \sum_{j=1}^{M+N} \alpha_j \Phi_j(x - x_j) \tag{9}$$

alakban, ahol $\Phi_1, \Phi_2, \ldots, \Phi_{M+N}$ adott radiális bázisfüggvények, az $\alpha_1, \alpha_2, \ldots, \alpha_{M+N}$ együtthatók egyelőre ismeretlenek. Feltéve, hogy nemcsak (9) jobb oldalán álló kifejezés approximálja jól u-t, hanem annak Laplace-értékei is Δu-t, az ismeretlen együtthatókra az alábbi rendszert nyerjük:

$$\sum_{j=1}^{M+N} \alpha_j \Delta\Phi_j(x - x_j) = f(x_k) \qquad (k = 1, 2, \ldots, M),$$

$$\sum_{j=1}^{M+N} \alpha_j \Phi_j(x - x_j) = u_0(x_k) \qquad (k = M+1, M+2, \ldots, M+N).$$

Ez az egyenletrendszer – nagy M és N esetén – épp olyan rossz numerikus tulajdonságokkal rendelkezik, mint a korábban bemutatott RBF-interpolációs módszerekkel kapcsolatos egyenletrendszer: a mátrix nagyméretű, általában teljesen kitöltött, sokszor nemszimmetrikus és rosszul kondicionált.

2. *A partikuláris megoldások módszere*: Ekkor először (8) jobb oldalán álló f függvényt approximáljuk:

$$f(x) := \sum_{j=1}^{M+N} \alpha_j \Phi_j(x - x_j) \tag{10}$$

alakban, ahol $\Phi_1, \Phi_2, \ldots, \Phi_{M+N}$ adott radiális bázisfüggvények, az $\alpha_1, \alpha_2, \ldots, \alpha_{M+N}$ együtthatók egyelőre ismeretlenek. Megoldva a megfelelő

$$\sum_{j=1}^{M+N} \alpha_j \Phi_j(x_k - x_j) = f(x_k) \qquad (k = 1, 2, \ldots, M+N) \tag{11}$$

interpolációs egyenletrendszert, tekintsük az ugyanezen $\alpha_1, \alpha_2, \ldots, \alpha_{M+N}$ együtthatókkal képzett v függvényt:

$$v(x) := \sum_{j=1}^{M+N} \alpha_j \Psi_j(x - x_j),$$

ahol Ψ_j-k olyan radiális bázisfüggvények, melyekre $\Delta\Psi_j = \Phi_j$ teljesül. (Ilyen Ψ_j függvények általában analitikusan megadhatók Φ_j-k ismeretében.) Akkor v (közelítő) megoldása az (8) Poisson-egyenletnek, mert:

$$\Delta v(x) := \sum_{j=1}^{M+N} \alpha_j \Delta\Psi_j(x - x_j) = \sum_{j=1}^{M+N} \alpha_j \Phi_j(x - x_j) = f(x).$$

Ennélfogva (8) u megoldása $u = v + w$ alakban előáll, ahol w megoldása a $\Delta w = 0$ Laplace-egyenletnek, és kielégíti a módosított

$$w|_{\partial\Omega} = u_0 - v|_{\partial\Omega}$$

peremfeltételt.

A problémát így egy *Laplace*-egyenlet Dirichlet-feladatára vezettük vissza, amely tartományon definiált függvényt már nem tartalmaz.

Megjegyezzük, hogy mind a Kansa-módszer, mind a partikuláris megoldások módszere nehézség nélkül alkalmazható általánosabb peremfeltételek mellett is.

Numerikus szempontból a partikuláris megoldások módszere sokszor előnyösebb lehet: a homogén egyenletre ui. esetleg speciális módszerek (pl. perem-integrálegyenlet módszer) alkalmazhatók. Az így nyert diszkretizált egyenletrendszer azonban általában a már említett rossz tulajdonságokkal rendelkezik.

3.3. Lokális sémák generálása a radiális bázisfüggvények módszerének alapján

A korábban említett numerikus hátrányok megkerülésének egyik lehetséges módja az, hogy a radiális bázisfüggvény-módszer helyett annak egy implicit változatát használjuk, amikor az interpolációs függvényt egy magasabbrendű parciális

differenciálegyenlet, pl. a biharmonikus egyenlet megoldásaként állítjuk elő [3]. Egy másik lehetséges technika, hogy a radiális bázisfüggvényekkel történő interpolációt csak *lokálisan* használjuk, mindig csak a szükséges kiértékelési x helynek csak egy alkalmas környezetébe eső alappontokat felhasználva [4], [5].

Az ilyen módszerek bizonyos szempontból a jól ismert véges differencia sémák általánosításainak tekinthetők. A sémák nem szabályos stencilen, hanem szabálytalanul elszórt pontokban lesznek definiálva. A lokális sémák jellegzetessége, hogy egy-egy pontban a vizsgált differenciáloperátort a *szomszédos* pontokhoz tartozó függvényértékekkel approximáljuk. Szomszédos pontokként elvben a legegyszerűbb az adott ponthoz egy adott távolságon belül elhelyezkedő pontokat tekinteni (mely távolság nem feltétlen ugyanaz minden pont esetén). Egy másik lehetőség: adott számú, a szóban forgó ponthoz legközelebbi pontok összessége. A továbbiakban feltesszük, hogy minden alappont szomszédai már definiálva vannak.

Modellfeladatként tekintsük ismét a Dirichlet-peremfeltétellel ellátott Poisson-egyenletet:

$$\Delta u = f \quad \Omega\text{-ban} \qquad u|_{\partial\Omega} = u_0. \tag{12}$$

Legyen $S := \{x_1, x_2, \ldots, x_N\} \subset \bar{\Omega}$ interpolációs alappontok egy véges halmaza. Célunk a Laplace-operátor diszkretizálása ezen a ponthalmazon. Ha u egy $\bar{\Omega}$-n értelmezett folytonos függvény, jelölje a rövidség kedvéért $u_k := u(x_k)$ $(k = 1, 2, \ldots, N)$. Lokális sémák konstruálásakor a differenciáloperátort tetszőleges x_m pontbeli diszkretizációjához csak az itt és a szomszédos pontokban felvett függvényértékeket használjuk. A klasszikus Taylor-sorfejtésen alapuló technika szórt alappontrendszeren meglehetősen nehézkes, ezért a sémákat az RBF-interpoláció felhasználásával konstruáljuk.

Legyen $x_m \in S$ tetszőleges, rögzített, centrálisnak tekintett alappont. Legyenek x_m szomszédai $x_1^{(m)}, x_2^{(m)}, \ldots, x_{N_m}^{(m)} \in S$, melyek egymástól és x_m-től is különböznek. Jelöljön $\tilde{u}_m$ egy, az $x_1^{(m)}, x_2^{(m)}, \ldots, x_{N_m}^{(m)}$ alappontokra támaszkodó lokális interpolációs függvényt:

$$\tilde{u}^{(m)}(x) := \sum_{j=1}^{N_m} \alpha_j \Phi\left(x - x_j^{(m)}\right) + \sum_{j=1}^{M} a_j p_j(x),$$

ahol Φ adott radiális bázisfüggvény, $p_1, \ldots, p_M$ pedig adott (alacsony fokszámú) polinomok (jellemzően az $1, x, y, x^2, xy, y^2, \ldots$ formulákkal definiált kétváltozós alappolinomok közül választva). Az ismeretlen $\alpha_1, \ldots, \alpha_{N_m}, a_1, \ldots, a_M$ együtthatók az interpolációs és az ortogonalitási feltételekből számíthatók:

$$\sum_{j=1}^{N_m} \alpha_j \Phi\left(x_k^{(m)} - x_j^{(m)}\right) + \sum_{j=1}^{M} a_j p_j\left(x_k^{(m)}\right) = u^{(m)}(x_k) \qquad (k = 1, 2, \ldots N_m),$$

$$\sum_{j=1}^{N_m} \alpha_j p_k\left(x_j^{(m)}\right) = 0 \qquad (k = 1, 2, \ldots M),$$

vagy tömören:

$$\begin{pmatrix} A & B \\ B^* & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \mathrm{a} \end{pmatrix} = \begin{pmatrix} \mathbf{u} \\ \mathbf{0} \end{pmatrix}, \tag{13}$$

ahol A szimmetrikus mátrix, $A_{kj} = \Phi\left(x_k^{(m)} - x_j^{(m)}\right)$ $(k,j = 1,\ldots,N_m)$, $B_{kj} = p_j\left(x_k^{(m)}\right)$ $(k = 1,\ldots,N_m,\ j = 1,\ldots,M)$, $\alpha := (\alpha_1,\ldots,\alpha_{N_m}) \in \mathbf{R}^{N_m}$, $\mathrm{a} := (a_1,\ldots,a_M) \in \mathbf{R}^M$. Az α_j és az a_j együtthatók esetén az m indexet az egyszerűség kedvéért nem tüntettük fel.

A (13) egyenletrendszer megoldhatósága általában nem biztosított, még akkor sem, ha történetesen A pozitív definit. Ekkor ui. (13) egyértelmű megoldhatósága könnyen láthatóan azzal ekvivalens, hogy B magtere csak a zérusvektorból áll, azaz B mint lineáris operátor, kölcsönösen egyértelmű. Ez pedig B elemeinek definíciója miatt azzal ekvivalens, hogy a $\sum_{j=1}^{M} a_j p_j(x)$ polinomok közül csak a zéruspolinom tűnik el mindegyik $x_1^{(m)},\ldots,x_{N_m}^{(m)}$ alappontban. Ez az $x_1^{(m)},\ldots,x_{N_m}^{(m)}$ pontok elhelyezkedésére ró nehezen ellenőrizhető feltételeket.

Ha a (13) rendszer történetesen minden $x_m \in S$ centrális pont esetén megoldható, akkor a (12) Poisson-egyenlet az alábbi módon diszkretizálható. Helyezzünk el x_m körül, a fő koordinátairányokban x_m-től $h > 0$ távolságra négy fiktív pontot $\left(x_m^N, x_m^W, x_m^S, x_m^E\right)$, ahol h jelöli az $\|x_m - x_k^{(m)}\|$ $(k = 1,2,\ldots,N_m)$ távolságok valamilyen közepét (a későbbiekben négyzetes közepet alkalmaztunk). A (12) Poisson-egyenlet diszkretizált alakja ekkor (Seidel-iterációra alkalmas formába írva):

$$u_m := \frac{\tilde{u}^{(m)}\left(x_m^N\right) + \tilde{u}^{(m)}\left(x_m^W\right) + \tilde{u}^{(m)}\left(x_m^S\right) + \tilde{u}^{(m)}\left(x_m^E\right)}{4} - \frac{h^2 \cdot f(x_m)}{4}. \tag{14}$$

A jobb oldali első tört az $u_k^{(m)}$ számokkal kifejezhető. Valóban, a szóban forgó tört definíció szerint:

$$\sum_{j=1}^{N_m} \alpha_j \frac{\Phi(x_m^N - x_j^{(m)}) + \Phi(x_m^W - x_j^{(m)}) + \Phi(x_m^S - x_j^{(m)}) + \Phi(x_m^E - x_j^{(m)})}{4} +$$

$$+ \sum_{j=1}^{M} a_j \frac{p_j(x_m^N) + p_j(x_m^W) + p_j(x_m^S) + p_j(x_m^E)}{4} =:$$

$$=: \sum_{j=1}^{N_m} \alpha_j \beta_j + \sum_{j=1}^{M} a_j b_j =: \left\langle \begin{pmatrix} \alpha \\ \mathrm{a} \end{pmatrix}, \begin{pmatrix} \beta \\ \mathrm{b} \end{pmatrix} \right\rangle,$$

ahol

$$\beta_j := \frac{1}{4} \cdot \left(\Phi\left(x_m^N - x_j^{(m)}\right) + \Phi\left(x_m^W - x_j^{(m)}\right) + \Phi\left(x_m^S - x_j^{(m)}\right) + \Phi\left(x_m^E - x_j^{(m)}\right)\right)$$
$$(j = 1,2,\ldots,N_m),$$

és

$$b_j := \frac{1}{4} \cdot \left(p_j\left(x_m^N\right) + p_j\left(x_m^W\right) + p_j\left(x_m^S\right) + p_j\left(x_m^E\right)\right) \quad (j = 1, 2, \ldots, M).$$

Mivel pedig nyilván

$$\begin{pmatrix} \alpha \\ \mathbf{a} \end{pmatrix} = \begin{pmatrix} A & B \\ B^* & 0 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{u} \\ \mathbf{0} \end{pmatrix},$$

azért innen

$$\left\langle \begin{pmatrix} \alpha \\ \mathbf{a} \end{pmatrix}, \begin{pmatrix} \beta \\ \mathbf{b} \end{pmatrix} \right\rangle = \left\langle \begin{pmatrix} \mathbf{u} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} A & B \\ B^* & 0 \end{pmatrix}^{-1} \begin{pmatrix} \beta \\ \mathbf{b} \end{pmatrix} \right\rangle =: \left\langle \begin{pmatrix} \mathbf{u} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{w} \\ \mathbf{v} \end{pmatrix} \right\rangle,$$

ahol a $\mathbf{w}$, $\mathbf{v}$ vektorpár megoldása az alábbi *lokális egyenletrendszernek*:

$$\begin{pmatrix} A & B \\ B^* & 0 \end{pmatrix} \begin{pmatrix} \mathbf{w} \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} \beta \\ \mathbf{b} \end{pmatrix}. \tag{15}$$

A lokális séma konstrukciója tehát a következő algoritmussal történik. Minden $x_m \in S$ ponthoz:

- meghatározzuk a szomszédos pontokat;
- kiszámítjuk a β és a $\mathbf{b}$ vektorokat;
- megoldjuk a (15) lokális egyenletrendszert.

Az így kapott $\mathbf{w}$, $\mathbf{v}$ vektorpárból $\mathbf{v}$-t később már nem használjuk; $\mathbf{w}$ elemeivel pedig felírhatjuk a diszkretizált Poisson-egyenletet (Seidel-iterációs formában):

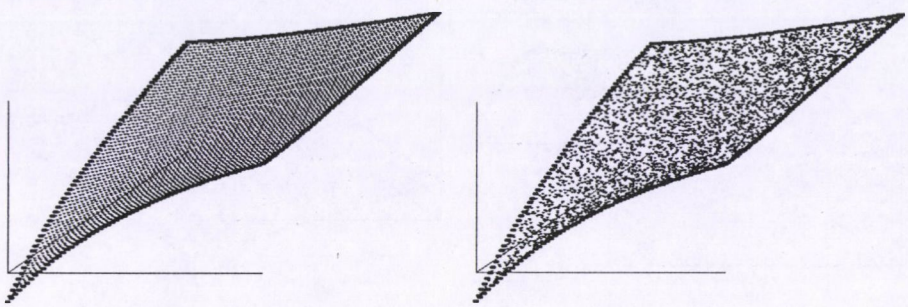
$$u_m := \sum_{j=1}^{N_m} w_j u_j^{(m)} - \frac{h^2 \cdot f(x_m)}{4}. \tag{16}$$

A w_j $(j = 1, 2, \ldots, N_m)$ együtthatókat elég a számítás elején egyszer meghatározni. A (16) Seidel-iterációt ezek után vagy önmagában használhatjuk mint a diszkretizált Poisson-egyenlet megoldási algoritmusát, vagy egy – eléggé természetes módon definiálható – multigrid környezetbe ágyazva, simító iterációként. Az iteráció során a peremfeltétel figyelembe vétele értelemszerűen, nehézség nélkül elvégezhető.

Példaként tekintsük az egységnégyzetben a Laplace-egyenletet. A pontos megoldás legyen $u(x, y) := \log\left((3x + 0,5)^2 + (3y + 0,5)^2\right)$, a Dirichlet-peremfeltétel pedig ezzel konzisztens. A tesztfeladat megoldására a fentebb leírt lokális sémát alkalmaztuk $\Phi(r) := r^2 \log r$ radiális bázisfüggvénnyel (polárkoordinátákban felírva), legfeljebb elsőfokú polinomokkal kiegészítve. A számításokat kétféle ponthalmazon hajtottuk végre: egy ekvidisztáns rácson és egy egyenletes eloszlás szerint kvázi-véletlenszerűen elszórt ponthalmazon. Az iteráció indításakor mindig zérus

N (M)	256 (64)	1024 (128)	4096 (256)
Relatív L_2-hiba (ekvidisztáns rács), %	0,1157	0,0285	0,0067
Relatív L_2-hiba (szórt ponthalmaz), %	0,1798	0,0480	0,0165

1. táblázat. Lokális séma alkalmazása a Laplace-egyenlet megoldására. A közelítő megoldások relatív L_2-hibái.



1. ábra. A Laplace-egyenlet közelítő megoldása lokális sémával. Tesztfeladat: $u(x,y) := \log((3x+0,5)^2 + (3y+0,5)^2)$.

kezdeti közelítésből indultunk ki. Három különböző alappontszám esetén a közelítések relatív L_2-hibáit az 1. táblázat mutatja (N jelöli a tartomány belsejében, M pedig a peremen elhelyezett pontok számát.) Az 1. ábrán pedig a közelítő megoldások láthatók $N = 4096$, $M = 256$ esetében. A számítások asztali számítógéppel készültek, a processzor órafrekvenciája 1,7 GHz volt. A két alacsonyabb pontszámú esetben (256, ill. 1024 belső pont) a relatív L_2-hiba 1 másodpercnél kevesebb idő alatt csökkent 1% alá, 4096 belső pont esetén kb. 9 másodperc alatt. Érdemes megemlíteni, hogy a tapasztalati konvergenciasebesség sokkal nagyobb volt, mint a hagyományos 5-pontos differenciasémák esetében. 1024 belső pont esetében pl. ekvidisztáns rácson 140, szórt pontrendszer esetében kb. 150 iterációs lépés alatt csökkent 1% alá a relatív L_2-hiba. Ez nem meglepő, mert minden egyes ponthoz tartozó lokális sémában a figyelembe vett szomszédos pontok száma itt jóval nagyobb, jellemzően 6 és 30 között változik: ennek megfelelően viszont az iteráció fajlagos számításigénye is nagyobb.

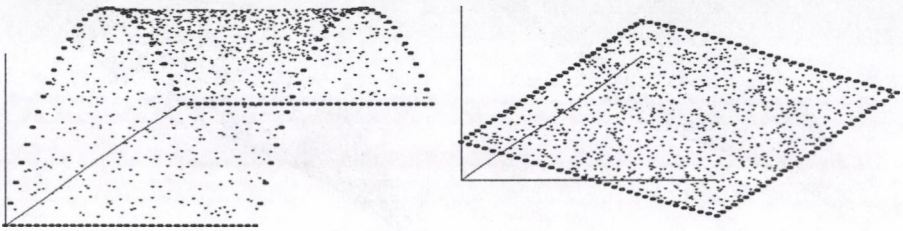
4. Numerikus eredmények

A fentebb leírt Stokes-egyenletmegoldó módszert illusztrálandó, tekintsük az alábbi tesztfeladatot. Legyen Ω az egységnégyzet, melyben vízszintes áramlást

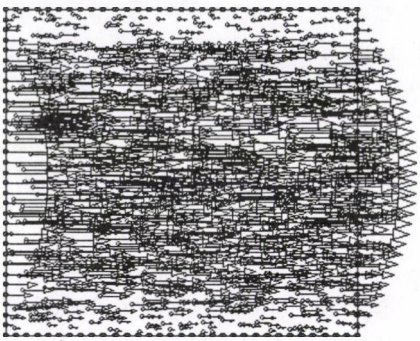
tételezünk fel kvadratikus sebességprofillal. Ekkor a pontos megoldás:

$$u(x,y) = c \cdot y(L-y), \qquad v(x,y) = 0, \qquad p(x,y) = c\nu\rho \cdot (L-2x),$$

ahol $L = 1$, c pedig egy $\frac{1}{\mathrm{m \cdot sec}}$ dimenziójú skálázó konstans. A tesztfeladatban a legegyszerűbb $c := 1$, $\nu\rho := 1$ választással éltünk. A tartomány peremét 128 ponttal diszkretizáltuk, a tartomány belsejében 1024 kvázi-véletlenül elszórt pontot helyeztünk el. A Poisson-egyenleteket lokális sémákkal diszkretizáltuk. A számított u vízszintes sebességkomponens és a p nyomás a 2. ábrán, a számított sebességmező pedig a 3. ábrán látható. A sebességmező relatív L_2-hibája 0.754% volt.



2. ábra. Számított vízszintes sebességkomponensek és nyomások.



3. ábra. Számított sebességek.

Köszönetnyilvánítás: A szerző köszönetét fejezi ki az Országos Tudományos Kutatási Alapprogramok szervezetének a kutatás részbeni finanszírozásáért. A téma nyilvántartási száma: T47287.

Hivatkozások

[1] FORTIN, M.: *Finite element solution of the Navier-Stokes equations.* Acta Numerica (1993), pp. 239–284.

[2] FRANKE, R.: *Scattered Data Interpolation: Test of Some Methods.* Mathematics of Computation, Vol. **38**, No. **157**, 1982.

[3] GÁSPÁR C.: *Fast multi-level meshless methods based on the implicit use of radial basis functions.* Lecture Notes in Computational Science and Engineering (2002) Vol. **26**, pp. 143–160, Springer-Verlag, Berlin, Heidelberg, New York.

[4] GÁSPÁR C.: *Global and Local Multi-level Meshless Schemes Based on Multi-Elliptic Interpolation.* Proceedings of ECCOMAS Thematic Conference on Meshless Methods, held in Lisbon, Portugal, July 11–14, 2005. (ed. by V.M.A.Leitao, C.J.S.Alves, C.A.Duarte), pp. B12.1–B12.6.

[5] GÁSPÁR C.: *Meshless Boundary Interpolation: Local and Global Multi-Level Techniques.* Advances in Boundary Element Techniques VII. Proceedings of the International Conference on Boundary Element Techniques VII, held in Paris, France, September 4–6, 2006. (ed. by B.Gatmiri, A.Sellers, M.H.Aliabadi), pp. 73–78.

[6] GOLBERG, M.A., CHEN, C.S.: *The theory of radial basis functions applied to the BEM for inhomogeneous partial differential equations.* Boundary Element Communications, 1994, 5(2), 57–61.

[7] GOLBERG, M.A., CHEN, C.S.: *A bibliography on radial basis function approximation.* Boundary Element Communications, 1996, 7(4), 155–163.

[8] HARDY, R.L.: *Theory and Applications of the Multiquadric-Biharmonic Method, 20 Years of Discovery 1968–1988.* Computers Math. Applic., Vol **19**, No. **8–9**, pp. 163–208, 1990.

[9] KANSA, E.J.: *Multiquadrics–a Scattered Data Approximation Scheme with Applications to Computational Fluid Dynamics–I. Surface Approximations and Partial Derivative Estimates.* Computers Math. Applic,Vol. **19**, No. **8/9**, pp. 127–145, 1990.

[10] KANSA, E.J.: *Multiquadrics–a Scattered Data Approximation Scheme with Applications to Computational Fluid Dynamics–II. Solutions to Parabolic, Hyperbolic and Elliptic Partial Differential Equations.* Computers Math. Applic., Vol. **19.**, No. **8/9**, pp. 147–161, 1990.

[11] MADYCH, W.R., NELSON, S.A.: *Multivariate interpolation and conditionally positive definite functions.* II. Math. Comput., Vol. **54**, pp. 211–230, 1990.

[12] MICCHELLI, C.A.: *Interpolation of scattered data: distance matrices and conditionally positive definite functions.* Const. Approx., Vol. **2**, pp. 11–22, 1986.

[13] PARTRIDGE, P.W., BREBBIA, C.A.: *Computer Implementation of the BEM Dual Reciprocity Method for the Solution of general Field Equations.* Communications in Applied Numerical Methods, 1990, **6**, 83–92.

[14] SIVALOGANATHAN, S., SHAW, G.J.: *A multigrid method for recirculating flows.* International Journal for Numerical Methods on Fluids, Vol. **8** (1988), pp. 417–440.

[15] STOYAN, G, TAKÓ G.: *Numerikus módszerek III.* ELTE-TypoTeX, Budapest, 1997.

GÁSPÁR CSABA
Széchenyi István Egyetem
H-9026 Győr, Egyetem tér 1.
gasparcs@sze.hu

MESHLESS METHODS WITH APPLICATION TO THE STOKES PROBLEM

CSABA GÁSPÁR

In this paper, a numerical solution technique of the two-dimensional permanent Stokes equations is presented. The proposed method is a meshless (meshfree) version of the pressure correction method based on the well-known Uzawa algorithm. The Poisson equations appearing in the pressure correction method are solved in a meshless way. Thus, the discretization of the flow domain (using either finite elements or a Cartesian or curvilinear grid) can be avoided. The applied schemes are based on the localised version of the Method of Radial Basis Functions. The computational complexity of the proposed method is relatively low, and the method can be embedded in a multi-level context in a natural way. The method is illustrated via a simple test flow problem with quadratic velocity profile.

Alkalmazott Matematikai Lapok **26** (2009), 223-254.

ELLIPSZIS PÁLYÁN MOZGÓ HENGER KÖRÜLI KIS REYNOLDS SZÁMÚ ÁRAMLÁS NUMERIKUS VIZSGÁLATA

BARANYI LÁSZLÓ

Ez a dolgozat a párhuzamos áramlásba helyezett, ellipszis pályán mozgó körhenger körül kialakuló kis Reynolds-számú összenyomhatalan folyadékáramlás kétdimenziós numerikus szimulációjával foglalkozik. A felhajtóerő-tényező, az ellenállástényező és a hátsó nyomástényező (base pressure coefficient) időátlagát és *rms* értékét, valamint a henger és a folyadék közötti energiacserére jellemző energiaátadási tényezőt a pálya ellipticitása függvényében ábrázolva, azok értékeiben ugrásszerű változások tapasztalhatók. Egy „ugrás" előtt és után határciklus analízist végeztünk, időbeli, fázisszög és áramkép változásokat vizsgáltunk. A vizsgálatok azt mutatják, hogy az ellipticitás nagy hatással lehet a mechanikusan mozgatott henger és folyadék közötti energiaátadásra, és hogy a transzverzális mozgás amplitúdójának kis mértékű megváltoztatása erősen befolyásolhatja az erőtényezőket. A felhajtóerő-tényező és a henger transzverzális elmozdulása közötti fázisszög az ugráson áthaladva 180°-ot változik. Ezeket a változásokat az okozhatja, hogy ennek a nemlineáris rendszernek valószínűleg két attraktora van, és hogy megoldás ezek melyikéhez vonzódik, az a probléma paramétereinek értékeitől függ.

1. Bevezetés

A levegő- vagy folyadékáramlásba helyezett nem áramvonalas, vagy más néven tompa testekről leváló örvények gyakran a szerkezet meghibásodását okozzák. Jó példa erre a Tacoma Narrows függőhíd (USA), amely örvényleválás által keltett csavarólengés miatt omlott össze 1940-ben. Egy másik eset a Japán-beli Monju atomerőműben történt, ahol az áramló folyadékba helyezett műanyag hőmérőtok a róla leváló örvények miatt kifáradt és megrepedt, a repedésen keresztül pedig primer hűtőfolyadék jutott ki a rendszerből. Az erőművet 1995-ös leállítása óta nem indították újra. A szélnek kitett magas karcsú épületekről, silókról, gyárkéményekről leváló örvények az építmény nagy amplitúdójú rezgéséhez vezethetnek, ha annak sajátfrekvenciája közel esik az örvényleválási frekvenciához és ugyanakkor a szerkezet csillapítása kicsi. A hőcserélőkben lévő csőkötegekről leváló örvények a hőcserélő rezgéséhez és kellemetlen zajos üzeméhez vezethetnek.

A körhenger körüli áramlást annak gyakorlati fontossága miatt igen sok kutató vizsgálja napjainkban is mind elméleti, mind kísérleti és numerikus eszközökkel.

A körhenger mögötti tér igen gazdag áramlástani jelenségekben: az örvényleválás szerkezete nagyon sok tényezőtől függ. Sok kutató foglalkozik a párhuzamos áramlásba helyezett álló és rezgő körhenger körüli áramlás vizsgálatával. Különösen sok tanulmány található álló henger esetére, lásd például [15], [31] és [39], de sokan foglalkoznak a henger transzverzális rezgőmozgásával, lásd [16] és [38], és a longitudinális irányban rezgő henger esetével is, lásd például [28]. A párhuzamos áramlásba helyezett elliptikus mozgást végző hengerek körüli esettel kevesebben foglalkoznak (lásd például [35]), annak ellenére, hogy ez a hullámokban mozgó henger körüli áramlás modelljének tekinthető.

A dolgozat következő fejezeteiben röviden ismertetünk egy eljárást az összenyomhatatlan newtoni közeg homogén párhuzamos áramlásába helyezett ellipszis pálya mentén keringő körhenger körüli két-dimenziós instacionárius kis Reynolds-számú áramlás számítására. A tanulmány az ellipszis pályán keringő henger esetén a mechanikai energiaátadási tényező, a fázisszög, a három erőtényező (felhajtóerő-tényező, ellenállás-tényező és hátsó nyomástényező) időátlagát és *rms* (root-mean-square) értékét vizsgálja egy eddig ismeretlen jelenséggel kapcsolatban: hirtelen ugrások lépnek fel az erőtényezők időátlagában és *rms* értékében, amikor a pálya ellipticitása függvényben rajzoljuk fel azokat.

Jelölésjegyzék

$\mathbf{a}_0$	a henger gyorsulásvektora, U^2/d-vel dimenziótlanítva
$A_{x,y}$	a rezgés amplitúdója x vagy y irányokban, d-vel dimenziótlanítva
C_D	ellenállás-tényező, $2F_D/\left(\rho U^2 d\right)$
C_L	felhajtóerő-tényező, $2F_L/\left(\rho U^2 d\right)$
C_{pb}	hátsó nyomástényező (base pressure coefficient), $\left\lfloor 2p/\left(\rho U^2\right)\right\rfloor_{\Theta=0}$
d	hengerátmérő, hosszlépték, $[m]$
D	sebesség divergenciája, U/d-vel dimenziótlanítva
e	ellipticitás, A_y/A_x
E	mechanikai energiaátadási tényező, $\rho U^2 d^2/2$-vel dimenziótlanítva
f	henger rezgési frekvenciája, U/d-vel dimenziótlanítva
$\mathbf{F}$	egységnyi hosszú hengerre ható erő, $\mathbf{F}_D\mathbf{i} + \mathbf{F}_L\mathbf{j}$
F_D	egységnyi hosszú hengerre ható ellenállás, [N/m]
F_L	egységnyi hosszú hengerre ható felhajtóerő, [N/m]
$\mathbf{i}$, $\mathbf{j}$	x, y irányú egységvektorok
p	folyadéknyomás, ρU^2-el dimenziótlanítva
R	sugár, d-vel dimenziótlanítva
Re	Reynolds-szám, Ud/ν
St	Strouhal-szám, örvényleválási frekvencia, U/d -vel dimenziótlanítva
t	idő, d/U -val dimenziótlanítva
T	örvényleválási periódus, d/U -val dimenziótlanítva
U	párhuzamos áramlás sebessége, sebességlépték, [m/s]

u, v x, y irányú sebességkomponensek, U-val dimenziótlanítva
$\mathbf{v}_0$ henger sebességvektora, U-val dimenziótlanítva
x, y Descartes-féle derékszögű koordináták, d-vel dimenziótlanítva
Φ fázisszög
ν kinematikai viszkozitási tényező, $[\mathrm{m}^2/\mathrm{s}]$
ρ folyadék sűrűsége, $[\mathrm{kg}/\mathrm{m}^3]$
Θ polárszög
ξ, η a számítási síkon lévő koordináták

Indexek

L felhajtóerő
D ellenállás
rms rms (root-mean-square) érték
x, y x vagy y irányú komponens
1, 2 energiaátadás y és x irányokban; a henger felületén, ill. a tartomány külső peremén
0 a hengermozgás elmozdulására, sebességére, gyorsulására

2. Alapegyenletek

Vizsgálatunk során egy összenyomhatatlan, állandó anyagjellemzőjű newtoni folyadék kétdimenziós (2D) áramlását tételezzük fel. Az alapegyenleteink a Navier–Stokes-egyenletek két komponenséből és a kontinuitási egyenletből állnak, amelyek dimenziótlan alakjai a tetszőleges a_0 gyorsulással mozgó hengerhez kötött koordinátarendszerben a következő módon írhatók fel:

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} = -\frac{\partial p}{\partial x} + \frac{1}{\mathrm{Re}}\nabla^2 u - a_{0x}, \tag{1}$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} = -\frac{\partial p}{\partial y} + \frac{1}{\mathrm{Re}}\nabla^2 v - a_{0y}, \tag{2}$$

$$D = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0, \tag{3}$$

Bár az (1)–(3) egyenletek elvileg alkalmasak az u, v sebességek és a p nyomás meghatározására, de azok időbeli változásának pontos megoldási lehetőségét jelentősen megnehezíti az a tény, hogy a (3) kontinuitási egyenlet nem tartalmazza explicite az idő szerinti deriváltat. A probléma áthidalható, ha egy külön egyenletet származtatunk a nyomásra. Az (1) egyenlet x szerinti és a (2) egyenlet y szerinti deriváltjait összeadva (azaz képezve a Navier–Stokes-egyenlet divergenciáját), a D sebességdivergenciát tartalmazó tagok közül csak annak idő szerinti parciális

deriváltját meghagyva, némi átrendezés után adódik a nyomásra vonatkozó Poisson-egyenlet (lásd [23]):

$$\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2} = 2\left[\frac{\partial u}{\partial x}\frac{\partial v}{\partial y} - \frac{\partial u}{\partial y}\frac{\partial v}{\partial x}\right] - \frac{\partial D}{\partial t}. \tag{4}$$

Mivel a henger $\mathbf{a}_0$ gyorsulása csak az időtől függ, így annak divergenciája zérus, ezért természetesen nem jelenik meg a (4) egyenletben. A fenti egyenletben a D a (3) egyenlet alapján ugyan zérus, de a fenti leírásmódot véges differenciák módszerével együtt alkalmazva nem elégíti ki egzakt módon a kontinuitást. Ezért a numerikus hibahalmozódás és az instabilitás elkerülése érdekében célszerű a (4) egyenletben a D idő szerinti parciális deriváltját meghagyni (lásd [23]). Baranyi és Shirakashi [3] a véges differenciák módszerén alapuló numerikus vizsgálata megerősítette, hogy a (4) utolsó tagjának elhagyása drasztikus hatással van a megoldásra, ugyanakkor az egyenletből elhagyott egyéb, D-t tartalmazó tagok csak elhanyagolható mértékben befolyásolják a megoldást.

Ezekben az egyenletekben u, v az x és y irányú sebességkomponens, p a nehézségi erőtér potenciáljával kibővített nyomás, D a sebesség divergenciája, Re az U áramlási sebesség, a d hengerátmérő és a folyadék ν kinematikai viszkozitási tényezőjéből számítható Reynolds-szám. Bár a (3) egyenlet alapján a D divergencia elméletileg zérus, annak idő szerinti deriváltját mégis meghagyjuk a (4) egyenletben a numerikus hibahalmozódás elkerülése érdekében (lásd [23]).

Peremfeltételek és kezdeti feltételek:

A $\mathbf{v}$ sebességre és a p nyomásra vonatkozó peremfeltételek az R_1 sugarú körhenger felületén valamint a számítási tartomány külső peremét jellemző, a körhengerrel azonos középpontú R_2 sugarú kör mentén, az alábbi módon adhatók meg (lásd 1. ábra):

(R_1) *hengerfelület:*

$$u = v = 0 \tag{5}$$

$$\frac{\partial p}{\partial n} = \frac{1}{\mathrm{Re}}\nabla^2 v_n - a_{0n}. \tag{6}$$

(R_2) *külső perem:*

$$u = u_{pot} - u_0, \qquad v = v_{pot} - v_0, \tag{7}$$

$$\frac{\partial p}{\partial n} = \left(\frac{\partial p}{\partial n}\right)_{pot}. \tag{8}$$

Az (5) egyenletből látható, hogy a henger felületén az u, v sebességkomponensek eltűnnek, míg a p nyomásra az (1) és (2) egyenletek felhasználásával a (6) Neumann-típusú peremfeltételt származtatjuk. A (6) egyenletben az n index a görbe külső normálisa irányában vett komponensre utal. A henger felületén kialakuló nyomáseloszlás – valamint a test és a folyadék között fellépő erő – pontos meghatározásához szükséges (6) összefüggésben szerepel a sebességkomponensek

felületi normális irányú, henger felületén vett deriváltjai, amelyeket a Taylor-sor felhasználásával nyert harmadrendű „féloldalas" differenciaséma segítségével származtatunk. A hengertől távoli, zavartalan áramlást jellemző R_2 sugár mentén potenciáláramlást tételezünk fel. Erre utal a (7) és (8) egyenletekben szereplő "*pot*" index. Megjegyezzük, hogy a potenciáláramlás feltételezése a tartomány külső peremén jó közelítést jelent a henger mögötti vékony holttértől eltekintve. A külső perem igen messze van a hengertől, így nem meglepő az a számítási tapasztalat (lásd [3]), hogy e feltevés mindössze a holttér külső tartományhatára környezetében torzítja el kis mértékben a sebességteret. A henger $\mathbf{a}_0$ gyorsulása és $\mathbf{v}_0$ sebessége (eltekintve természetesen a két mennyiség közötti összefüggéstől) tetszőleges lehet.

A dimenziótlan sebesség- és nyomáseloszlásásra vonatkozó kezdeti feltételként a számítások során a körhenger körüli potenciáláramlás sebesség- és nyomáseloszlását használjuk, amelyek az $U\mathbf{v}_0 = U\left(v_{0x}\left(t\right)\mathbf{i} + v_{0y}\left(t\right)\mathbf{j}\right)$ sebességgel mozgó hengerhez kötött rendszerben a dimenziótlan x, y koordináták segítségével a következő alakban írhatók fel:

$$u\left(x,y,t=0\right) = 1 - \frac{R_1^2\left(x^2-y^2\right)}{\left(x^2+y^2\right)^2} - v_{0x}\left(t=0\right), \quad \text{ha } x^2+y^2 > R_1^2 \tag{9}$$

$$v\left(x,y,t=0\right) = -2\frac{R_1^2 xy}{\left(x^2+y^2\right)^2} - v_{0y}\left(t=0\right), \quad \text{ha } x^2+y^2 > R_1^2 \tag{10}$$

$$p\left(x,y,t=0\right) = p_\infty + \frac{R_1^2}{\left(x^2+y^2\right)^2}\left(x^2-y^2-\frac{R_1^2}{2}\right), \tag{11}$$

ahol $R_1 = 0,5$ a körhenger d átmérőjével dimenziótlanított sugara, p_∞ a hengertől távoli zavartalan áramlásban érvényes dimenziótlan nyomás. Az (5) peremfeltétel alapján a henger felületén a $t = 0$ időpontban is előírjuk az

$$u\left(x,y,t=0\right) = v\left(x,y,t=0\right) = 0, \quad \text{ha } x^2+y^2 = R_1^2 \tag{12}$$

feltételt is.

A számításokat a $t = 0$ időpillanatban a hirtelen mozgásba hozott hengerhez tartozó kezdeti feltételek esetére is elvégeztük, amelyek az alábbi módon adhatók meg:

$$u\left(x,y,t=0\right) = U - v_{0x}\left(t=0\right), \quad \text{ha } x^2+y^2 > R_1^2 \tag{13}$$

$$v\left(x,y,t=0\right) = -v_{0y}\left(t=0\right), \quad \text{ha } x^2+y^2 > R_1^2 \tag{14}$$

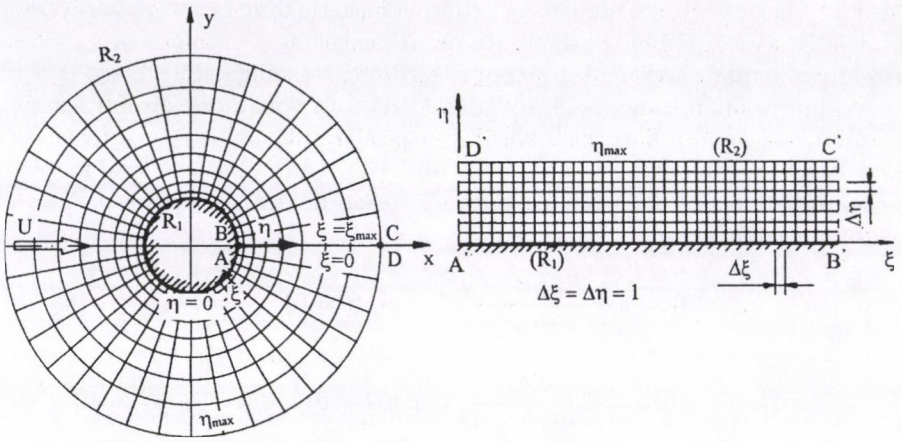
$$p\left(x,y,t=0\right) = p_\infty, \tag{15}$$

továbbá itt is előírjuk a (12) feltételt. A szerző (9)–(12) valamint a (13)–(15) kezdeti feltételek alapján nyert számítási eredményei elhanyagolható mértékben különböznek egymástól.

Természetesen az (1)–(4) alapegyenletek, (5)–(8) peremfeltételek és a fenti kezdeti feltételek álló hengerre is érvényesek maradnak az $\mathbf{a}_0 = \mathbf{0}$ és $\mathbf{v}_0 = \mathbf{0}$ helyettesítésekkel.

3. A tartomány és az alapegyenletek transzformációja

Azért, hogy a diszkretizációhoz használandó véges differenciák módszerének alkalmazása során a peremfeltételeket pontosan ki tudjuk elégíteni, és elkerülhessük a számítási pontosságot rontó interpolációt, peremre illesztett koordinátákat használunk. A fizikai sík számítási síkra való leképzését az 1. ábra mutatja.



1. ábra. A fizikai és számítási síkok

A fizikai sík (x, y) és a számítási sík (ξ, η) koordinátái közti kapcsolatot az

$$\begin{aligned} x(\xi, \eta) &= R(\eta) \cos[g(\xi)], \\ y(\xi, \eta) &= -R(\eta) \sin[g(\xi)] \end{aligned} \tag{16}$$

alakban vesszük fel, ahol

$$R(\eta) = R_1 \exp[f(\eta)]. \tag{17}$$

Az R_1 és R_2 sugarú hengerfelületeknek – ahol az (5)–(8) peremfeltételeket ki kell elégíteni – a számítási síkon az $\eta = 0$, ill. az $\eta = \eta_{\max}$ egyenesek felelnek meg. A ξ és η koordinátákat egész számokra választottuk, amelyek egyben a diszkretizációs pontok kerületi-, ill. sugárirányú sorszámainak vagy indexeinek is felfoghatók. Figyelembe véve még a (16) és (17) egyenleteket is, könnyen belátható, hogy a számítási síkon ortogonális egyenközű hálót nyerünk, amely azért is előnyös, mert a differenciasémák többsége erre az esetre van kidolgozva, és azok általában ilyenkor magasabb rendű közelítést jelentenek, mint nem egyenközű háló esetén.

Bár többféle $f(\eta)$ és $g(\xi)$ függvényt kipróbáltunk, ebben a dolgozatban az egyszerű

$$g(\xi) = 2\pi \frac{\xi}{\xi_{\max}}, \quad f(\eta) = \frac{\eta}{\eta_{\max}} \ln\left(\frac{R_2}{R_1}\right) \tag{18}$$

lineáris leképzőfüggvény alkalmazásával nyert eredményeket mutatjuk be. Az $f(\eta)$ és $g(\xi)$ függvények ilyen megválasztása is biztosítja, hogy a henger közelében – ahol a sebesség erősen változik – a háló sűrű, attól távolodva pedig egyre ritkább legyen. A (16)–(18) leképzés kölcsönösen egyértelmű, mert a J Jacobi-féle determináns

$$J = y_\eta x_\xi - y_\xi x_\eta = \frac{2\pi \ln(R_2/R_1)}{\xi_{\max}\eta_{\max}} R^2(\eta) \tag{19}$$

tetszőleges ξ és η értékekre pozitív értéket ad. A (19) egyenletben a ξ és η indexek differenciálást jelölnek.

A fizikai síkon (lásd 1. ábra) a görbe vonalú háló egy elemi négyszöge két oldalának a hányadosa – az ún. rácsviszony – [22] alapján a következő alakban írható fel:

$$\mathrm{AR} = \sqrt{\frac{g_{22}}{g_{11}}} = \frac{f_\eta}{g_\xi} = \frac{\xi_{\max}\ln(R_2/R_1)}{2\pi\eta_{\max}}, \tag{20}$$

ahol g_{11} és g_{22} a metrikus tenzor elemei, és az egyenletben a ξ és η indexek most is differenciálást jelölnek. A (20) egyenletből látható, hogy a rácsviszony a (18) lineáris leképzőfüggvények esetén az egész fizikai tartományon állandó. A két egymásra merőleges irányban vett rácspontok számának $(\xi_{\max}, \eta_{\max})$ és a vizsgált tartomány méretére jellemző R_2/R_1 sugárviszony megfelelő választásával elérhető, hogy az AR értéke pontosan 1 legyen, így megvalósítható az előnyös számítási tulajdonságokkal rendelkező konformis leképzés. A módszer előnye, hogy a (16)–(18) leképzés biztosítja, hogy a metrikus tenzor mellékátlóiban lévő elemek nullák legyenek, azaz $g_{12} = g_{21} = 0$. Így az (1), (2) és (4) egyenletekben szereplő Laplace-deriváltak transzformálásakor a vegyes másodrendű deriváltak eltűnnek. A (18) leképzőfüggvények lineáris volta miatt további egyszerűsödést jelent az, hogy a Laplace-deriváltak transzformálásakor az elsőrendű deriváltak is kiesnek, lásd [5]. Mivel a leképzés zárt alakban elemi függvényekkel van megadva, a metrikus paraméterek és a koordináta-deriváltak szintén zárt alakban számíthatók, így nincs szükség a számítási hibát okozó numerikus differenciálásra.

A (16)–(18) leképzés felhasználásával az (1)–(4) alapegyenleteket, az (5)–(8) peremfeltételeket és a (9)–(12) kezdeti feltételeket leképezzük a számítási síkra. Az (1) és (2) Navier–Stokes-egyenletek transzformált megfelelői a következők:

$$\frac{\partial u}{\partial t} + \frac{1}{J}\left(u\frac{\partial y}{\partial \eta} - v\frac{\partial x}{\partial \eta}\right)\frac{\partial u}{\partial \xi} + \frac{1}{J}\left(v\frac{\partial x}{\partial \xi} - u\frac{\partial y}{\partial \xi}\right)\frac{\partial u}{\partial \eta} =$$

$$-\frac{1}{J}\left(\frac{\partial y}{\partial \eta}\frac{\partial p}{\partial \xi} - \frac{\partial y}{\partial \xi}\frac{\partial p}{\partial \eta}\right) + \frac{1}{\mathrm{Re}J^2}\left(g_{22}\frac{\partial^2 u}{\partial \xi^2} + g_{11}\frac{\partial^2 u}{\partial \eta^2}\right) - a_{0x}, \tag{21}$$

$$\frac{\partial v}{\partial t} + \frac{1}{J}\left(u\frac{\partial y}{\partial \eta} - v\frac{\partial x}{\partial \eta}\right)\frac{\partial v}{\partial \xi} + \frac{1}{J}\left(v\frac{\partial x}{\partial \xi} - u\frac{\partial y}{\partial \xi}\right)\frac{\partial v}{\partial \eta} =$$

$$-\frac{1}{J}\left(\frac{\partial x}{\partial \xi}\frac{\partial p}{\partial \eta} - \frac{\partial x}{\partial \eta}\frac{\partial p}{\partial \xi}\right) + \frac{1}{\mathrm{Re}J^2}\left(g_{22}\frac{\partial^2 v}{\partial \xi^2} + g_{11}\frac{\partial^2 v}{\partial \eta^2}\right) - a_{0y}. \tag{22}$$

A D sebességdivergencia transzformációjára a következőt kapjuk:

$$D = \frac{1}{J}\left(\frac{\partial y}{\partial \eta}\frac{\partial u}{\partial \xi} - \frac{\partial y}{\partial \xi}\frac{\partial u}{\partial \eta} + \frac{\partial x}{\partial \xi}\frac{\partial v}{\partial \eta} - \frac{\partial x}{\partial \eta}\frac{\partial v}{\partial \xi}\right) = 0. \tag{23}$$

A nyomásra vonatkozó Poisson-egyenlet transzformáltja pedig a

$$g_{22}\frac{\partial^2 p}{\partial \xi^2} + g_{11}\frac{\partial^2 p}{\partial \eta^2} = 2J\left(\frac{\partial u}{\partial \xi}\frac{\partial v}{\partial \eta} - \frac{\partial u}{\partial \eta}\frac{\partial v}{\partial \xi}\right) - J^2\frac{\partial D}{\partial t} \tag{24}$$

alakot ölti.

Az egyenletekhez hasonlóan az (5)–(8) peremfeltételeket is leképezzük a számítási síkra. A (16)–(18) leképzés alkalmazása után a nyomásra vonatkozó (6) és (8) peremfeltételek a következő alakokba mennek át:

$$\text{ha } R = R_1: \qquad \frac{\partial p}{\partial \eta} = \frac{g_{11}}{\mathrm{Re}J^2}\left(\frac{\partial x}{\partial \eta}\frac{\partial^2 u}{\partial \eta^2} + \frac{\partial y}{\partial \eta}\frac{\partial^2 v}{\partial \eta^2}\right) - \frac{\partial x}{\partial \eta}a_{0x} - \frac{\partial y}{\partial \eta}a_{0y}, \tag{25}$$

$$\text{ha } R = R_2: \qquad \frac{\partial p}{\partial \eta} \cong \frac{R_2}{\eta_{\max}} ln\left(\frac{R_2}{R_1}\right)\left(\frac{\partial p}{\partial n}\right)_{pot}. \tag{26}$$

A sebességre vonatkozó peremfeltételek és a kezdeti feltételek transzformációjának bemutatásától eltekintünk. A (21), (22), (24) és (25) egyenletekben szereplő metrikus tenzor főátlójában lévő g_{11}, ill. g_{22} elemei a következő alakúak:

$$g_{11} = \left(\frac{\partial x}{\partial \xi}\right)^2 + \left(\frac{\partial y}{\partial \xi}\right)^2, \qquad g_{22} = \left(\frac{\partial x}{\partial \eta}\right)^2 + \left(\frac{\partial y}{\partial \eta}\right)^2.$$

A transzformáció (16), (17) alakú megválasztása tetszőleges $f(\eta)$ és $g(\xi)$ függvények esetén biztosítja, hogy a metrikus tenzor főátlón kívüli elemei nullák legyenek, azaz $g_{12} = g_{21} = 0$. Ezért hiányoznak a vegyes másodrendű deriváltak a (21), (22) és (24) egyenletekben szereplő Laplace-deriváltakból. A (18) leképzőfüggvények lineáris megválasztása egyben azt is biztosítja, hogy az előbb említett Laplace-deriváltak transzformált alakjaiból az elsőrendű deriváltak is kiessenek, (lásd [5]). Mivel az $f(\eta)$ és $g(\xi)$ leképzőfüggvények elemi függvényekkel adottak, a számításhoz szükséges metrikus paraméterek és a koordináta-deriváltak zárt alakban származtathatók, így nincs szükség a számítási hibához vezető numerikus differenciálásra. Mint már említettük, mind a számítási síkon nyert egyenközű háló, mind a fizikai síkon az AR rácsviszony egységnyire választásával nyerhető konformis leképzés előnyös számítástechnikai szempontból. További előny, hogy mivel a számítási hálót a mozgó hengerhez rögzítjük, és a számítások során nem változtatjuk, ezért nincs szükség minden egyes időlépcsőben új hálót létrehozni, elég a hálógenerálást egyszer, a számítások előtt elvégezni. Ennek további előnye az, hogy a transzformációs egyenleteink egyszerűbbek, mivel nem szerepelnek benne a háló deformációjára jellemző tagok. Látható tehát, hogy ez a számítási eljárás több szempontból is optimalizált.

4. Numerikus módszer és számítási eredmények

Az egyenletek megoldására a véges differenciák módszerét használjuk. A térbeli deriváltakat a tartomány belsejében negyedrendű centrális differenciákkal közelítjük, a tartomány szélei közelében ugyancsak negyedrendű, de nem centrális differenciákat használunk, a konvektív deriváltak közelítésére a [24] által kidolgozott harmadrendű módosított *upwind* sémát alkalmazzuk. Az időbeli diszkretizációhoz elsőrendű haladó differenciákat használunk. A módszer részletesebb leírása a [3] és [5] dolgozatokban található. A sebességet a (21), (22) egyenletek közvetlen integrálásával kapjuk, miközben minden időlépcsőben kielégítjük a (23) kontinuitási egyenletet. A nyomáseloszlást a (24) Poisson-egyenletből a szukszcesszív túlrelaxálás módszerének segítségével határozzuk meg minden időlépcsőben. A relaxációs paraméter értékét 1,8-ra választva, viszonylag gyors konvergenciát tapasztaltunk. A relatív hibakorlát 10^{-5}-re, ill. 10^{-6}-ra választása gyakorlatilag elhanyagolható különbséget jelentett a megoldásban. Ugyan a szerző tudatában van annak, hogy ennél modernebb és hatékonyabb számítási eljárások is léteznek, de mivel a jelen módszer alkalmazásával a szakirodalomban található kísérleti eredményekkel jól egyező eredményeket kapott elfogadható számítási idő mellett, így nem tekintette elsődleges szempontnak a modernebb eljárás alkalmazását. A numerikus diszkretizáció során a (24) egyenletben a D divergencia n. időlépcsőben érvényes idő szerinti parciális deriváltját a

$$\frac{\partial D^n}{\partial t} \cong \frac{D^{n+1} - D^n}{\Delta t} = -\frac{D^n}{\Delta t} \tag{27}$$

összefüggéssel közelítjük, amely során a divergencia D^{n+1} új értékét nullának választjuk. Az elvégzett numerikus vizsgálataink azt mutatják, hogy ily módon a D^n divergencia térben és időben is egy igen kis érték alatt tartható (amely, mint tudjuk, a (3) és (23) egyenletek alapján elméletileg nulla).

A körhenger felületén lévő nyomáseloszlást – amely rendkívül fontos a körhenger és folyadék között fellépő erő meghatározásához – a következő harmadrendű formula segítségével származtattuk a Taylor-sor felhasználásával, figyelembe véve, hogy $\Delta\eta = 1.0$

$$p_{i,1} = \frac{-p_{i,3} + 4p_{i,2} - \left[2\frac{\partial p}{\partial \eta}\right]_{i,1}}{3}. \tag{28}$$

A (28) egyenletben az első (i) index a kerületi, a második index pedig a sugárirányú koordinátára utal (a második index 1 értéke a körhenger felületén lévő jellemzőkre vonatkozik). Az egyenletben szerepel a nyomás felületi normális irányában vett (η szerinti) deriváltja is, amely a (6), ill. (25) peremfeltétellel van megadva. Ezt a nyomásderiváltat minden időlépcsőben a körhenger környezetében lévő sebességeloszlás és a henger a_0 gyorsulásának felhasználásával határozzuk meg. A (25) egyenletben a g_{11} metrikus tenzorelem, és az x és y koordináták ξ és η szerinti deriváltjai analitikus alakban származtathatók, valamint a henger gyorsulás a_{0x} és a_{0y} komponenseinek időbeli változása is ismert. Így csak az u és v sebességkomponensek fal menti η szerinti második deriváltja igényel komolyabb figyelmet,

amelyre a Taylor-sor alkalmazásával kapjuk a következő harmadrendű formulát:

$$\left[\frac{\partial^2 u}{\partial \eta^2}\right]_{i,1} = \frac{35u_{i,1} - 104u_{i,2} + 114u_{i,3} - 56u_{i,4} + 11u_{i,5}}{12\,(\Delta\eta)^2}. \tag{29}$$

Az (5) peremfeltétel alapján a falon a sebesség eltűnik, így $u_{i,1} = 0$, másrészt $\Delta\eta = 1.0$, így a (29) összefüggés tovább egyszerűsödik.

Az alternáló jelek spektrumát a gyors Fourier-transzformációval (FFT) kapjuk, amelyekből meghatározható az örvényleválás frekvenciája.

A következőkben bemutatandó néhány számítási eredmény többsége esetében $R_2/R_1 = 40$; a dimenziótlan időlépcső: Δt =0,0005; rácspontok száma: 301x177, de kipróbáltuk a Δt =0,00025 és 481x288 értékeket is. Tapasztalatunk az, hogy ilyen háló esetén a megoldások gyakorlatilag már „hálófüggetlen"-nek tekinthetők. A 2. ábra (lásd a [13] dolgozatot is) egy, a főáramláshoz képest merőleges irányban rezgő hengerre (Re = 185; $Ax = 0$; $Ay = 0,2$; $f/\mathrm{St}_0 = 0,8$; $\mathrm{St}_0 = 0,195$) vonatkozó, a tehetetlenségi erőt nem tartalmazó (lásd [8]) felhajtóerő-tényező időbeli változását mutatja a fent említett két háló és időlépcső kombinációra. A pontvonal a durvább (301x177; $\Delta t = 0,0005$), a folytonos vonal a finomabb (481x288; $\Delta t = 0,00025$) hálóhoz tartozik. Mivel a két megoldás gyakorlatilag egybeesik, ezért a számítások többségét a kisebb számítási időt igénylő durvább hálón végeztük.

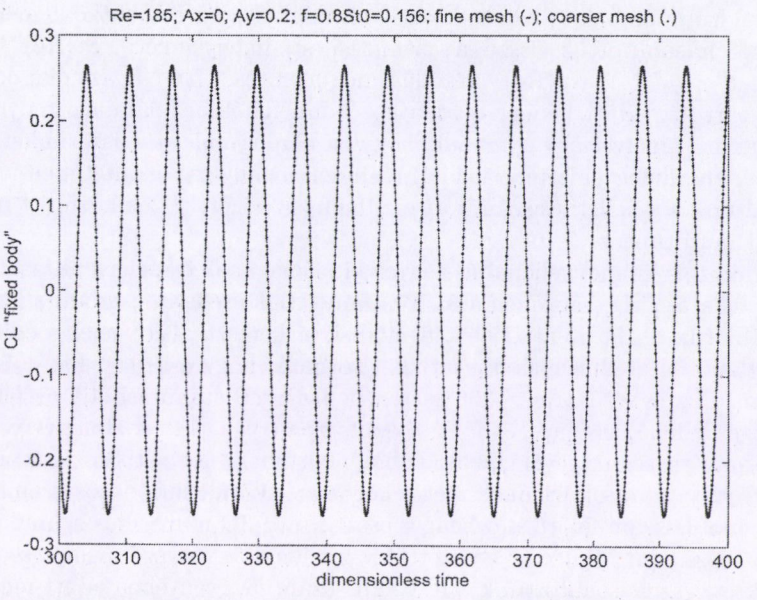
Megjegyezzük, hogy a rácspontok számát úgy választottuk meg, hogy a (20) egyenlettel definiált AR rácsviszony közelítőleg egységnyi legyen (amely a R_2/R_1 sugárviszony változtatásával pontosan egységnyi értékűre is beállítható). A nyert sebesség- és nyomáseloszlás ismeretében számos más jellemző is kiszámítható: például az áram- és örvényvonal eloszlás, fejhajtóerő- és ellenállás-tényező, nyomástényező, nyomatéktényező időbeli változása, a torlópont és a leválási pontok időbeli vándorlása és egyéb jellemzők.

Egy tetszőleges periodikus f függvény $\bar{f}$ időátlagát és f_{rms} *rms* (root-mean-square) értékét az

$$\bar{f} = \frac{1}{mT}\int_{t_1}^{t_1+mT} f\,(t)\,dt; \qquad f_{rms} = \sqrt{\frac{1}{mT}\int_{t_1}^{t_1+mT} \left[f(t) - \bar{f}\right]^2 dt}$$

összefüggésekből numerikus integrálás felhasználásával számítottuk, ahol t_1 az integrálás alsó határa, T egy örvényleválási ciklus (amely során a henger felső és alsó felületén is leválik egy-egy örvény) és m a számításhoz alapul vett ciklusok száma. Mind az *rms* értékeket, mind az időátlagokat több m érték esetére meghatároztuk a pontosság fokozása céljából. A periodikussá vált erőtényezők időátlagát és *rms* értékét általában 80-100 örvényleválási ciklus alapján határoztuk meg, de végeztünk 2000 periódusra vonatkozó számításokat is. A hosszú és rövid számításokon alapuló eredmények átlaga és *rms* értéke gyakorlatilag egybeesett.

Mint ismeretes (lásd pl. [2], [22]), az explicit módszer esetén az időlépcsőt elég kicsire kell választani ahhoz, Courant–Friedrichs–Levy (CFL) stabilitási feltétel teljesüljön. Mivel a nagy pontosság eléréséhez egyébként is kis időlépcsőre van



2. ábra. A tehetetlenségi erőtől mentes felhajtóerő-tényező időbeli változása két számítási háló és időlépcső kombinációra (pontvonal: 301x177, $\Delta = 0,0005$; folytonos vonal: 481x283, $\Delta t = 0,00025$) a főáramlásra merőleges irányban rezgő henger esetén (Re $= 185$, $A_x = 0$; $A_y = 0,2$, $f = 0,8\mathrm{St}_0$, $\mathrm{St}_0 = 0,195$)

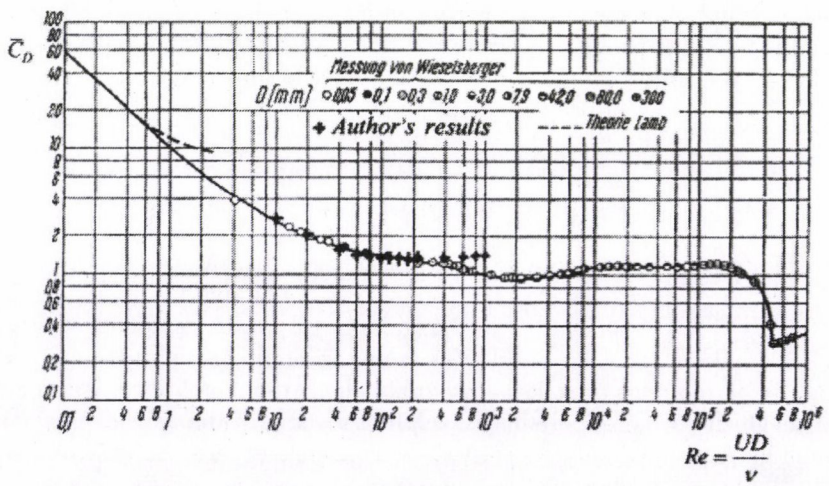
szükség, számításaink során mindössze igen kis Reynolds-számoknál (Re ≤ 5) fordult elő instabilitási probléma, amely az időlépcső további csökkentésével mindig elkerülhető volt. Mivel a jelen dolgozat az örvényleválásos tartományra koncentrál (Re > 47), így az adott vizsgálatok során CFL stabilitási feltétel semmilyen korlátozást nem jelentett.

4.1. Összehasonlítás mérési és számítási eredményekkel

E probléma megoldására a szerző által kidolgozott számítógépes eljárás hibakorlátjának meghatározása igen bonyolult lenne. E helyett a számítási háló és az időlépcső változtatása hatásának vizsgálatán túl fontos eszköz lehet a szakirodalomban rendelkezésre álló kísérleti és számítási eredményekkel történő összehasonlítás. A párhuzamos áramlásba helyezett álló körhengerekre vonatkozóan nagy számú megbízható mérési eredmény található a szakirodalomban, amely sajnos nem mondható el a gyorsuló mozgást végző hengerek esetére. Ezért stratégiánk az, hogy az álló hengerre vonatkozó számítási eredmények kísérleti eredményekkel való egyezése után terjesztjük ki a számítási eljárást a bonyolultabb, gyorsuló mozgást végző henger esetére.

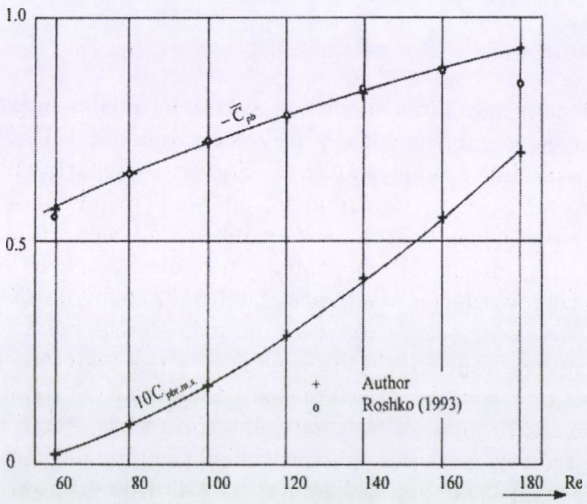
A szakirodalomban található következő mennyiségekre vonatkozó mérési eredményekkel hasonlítottuk össze a számítási eredményeinket: St(Re), $\bar{C}_D$(Re), $\bar{C}_{pb}$(Re), C_{Lrms}(Re), Nu(Re), áramlás megjelenítés. Itt Nu az állandó hőmérsékleten tartott, fűtött hengerre vonatkozó dimenziótlan hőátadási tényező, vagy Nusselt-szám, amelyet úgy nyertünk, hogy az alapegyenlet-rendszerünket kiegészítettük egy energiaegyenlettel (lásd [5]). Minden mennyiség tekintetében jó egyezést tapasztaltunk a kísérleti és számított eredmények között. Ezek közül itt most csak néhányat mutatunk be.

A számított és mért ellenállás-tényező időátlagának összehasonlítását mutatja a Reynolds-szám függvényében a 3. ábra, amelynek eredetije például a [34] könyvben található. A kör alakú jelek a mérési, a kereszt alakú jelek pedig a jelen szerző számítási eredményeit mutatják. Látható, hogy az $10 < \mathrm{Re} < 200$ tartományban jó az egyezés, $\mathrm{Re} > 200$ esetén pedig a 2D számítási eljárás felülbecsüli a tényleges ellenállás-tényezőt. Ez összhangban van [14] eredményeivel, akik a Floquet-analízis segítségével igazolták, hogy körülbelül $\mathrm{Re} = 190$-nél három-dimenziós (3D) instabilitások jelennek meg a körhenger körüli súrlódásos áramlásban, és így e Reynolds-szám fölött álló henger esetén már 3D numerikus szimulációs eljárást kell használni. Mint már említettük, az ellenállás-tényező mellett számos más jellemzőt is összehasonlítottunk a Reynolds-szám függvényében adott mérési eredményekkel, így a hátsó nyomástényező időátlagát, a dimenziótlan örvényleválási frekvenciát vagy Strouhal-számot és a felhajtóerő tényező rms értékét. Ezeken túl, az alapegyenleteket egy energiaegyenlettel kiegészítve, meghatároztuk a párhuzamos áramlásba helyezett álló, állandó felületi hőmérsékletű körhenger dimenziótlan hőátadási tényezőjét, az ún. Nusselt-számot, amelynek időátlagát szintén össze



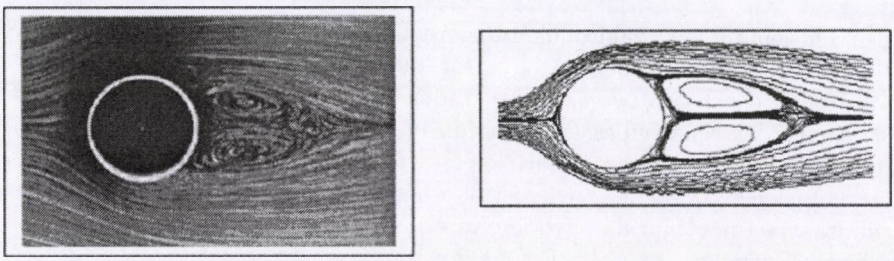
3. ábra. Az ellenállás-tényező időátlaga a Reynolds-szám függvényében

tudtuk hasonlítani a szakirodalomban rendelkezésre álló mérési eredményekkel. Az összes felsorolt esetben jó egyezést találtunk. A hely hiánya miatt ezek közül csak a C_{pb} hátsó nyomástényező időátlagának kísérleti eredményekkel (lásd [33]) történő összehasonlítását mutatjuk be (lásd 4. ábra). Az ábrán jó egyezés látható, eltekintve az $\mathrm{Re} = 180$ értékhez tartozó C_{pb} értékétől. Ebben a pontban a mérésnél jelenlevő zavarások felerősítik a három-dimenziós hatásokat; valószínűleg ez okozhatja e pontban a nagyobb eltérést. Az ábra tartalmazza még a C_{pb} rms értékének Reynolds-számtól való függését is.



4. ábra. A C_{pb} időátlaga és rms értéke a Reynolds-szám függvényében

Mint például a [36] dolgozatból ismeretes, a párhuzamos áramlásba helyezett álló körhenger esetén 47 alatti Reynolds-számoknál az áramlás stacionárius, és a körhenger mögött két álló ikerörvény helyezkedik el, amelyek nem válnak le a hengerről. Az $\mathrm{Re} = 47$ környezetében a Hopf-bifurkáció miatt megindul a henger két oldalán a periodikus örvényleválás, és ennél nagyobb Re érték esetén az áramlás instacionáriussá válik. Sok szerző végzett mind a stacionárius, mind az instacionárius áramlási tartományban áramlás-megjelenítéses kísérleteket; az eredmények gazdag tárháza a [20] könyv. A jelen szerző szerzőtársával (lásd [37]) számos körhenger körüli stacionárius áramlásra kiszámította az áramképeket, és összehasonlítva azokat áramlás-megjelenítéses fotókkal, igen jó egyezést tapasztalt. Itt a helyhiány miatt csak egy ilyen esetet mutatunk be (lásd a 5. ábra). A vizsgált esetben $\mathrm{Re} = 26$. A bal oldali ábra az alumínium porral történő kísérleti áramlás-megjelenítést mutatja, amelyet Taneda készített 1956-ban és a [20] könyvben található, a jobb oldalon pedig a számított áramkép található. Látható, hogy az egyezés igen jó.

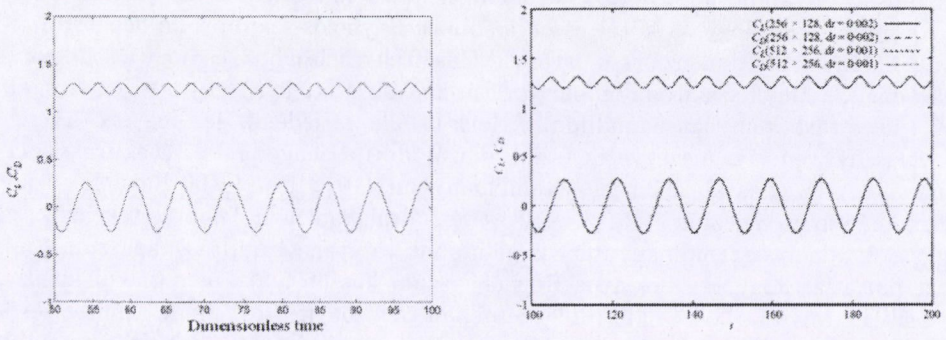


5. ábra. Körhenger körüli áramlás Re = 26 esetén. Bal oldal: kísérleti; jobb oldal: számított

A kísérleti összehasonlítás mellett a számítási eredményeket más számítási módszerekkel nyert eredményekkel is összehasonlítottuk. Az egyik ilyen volt a spektrális elem módszer, amely során S. Sherwin (Department of Aeronautics, Imperial College of Science, Technology and Medicine, London, UK) nem publikált eredményeivel hasonlítottuk össze saját számításainkat, amelyről a [4]-ben számoltunk be. Egy másik összehasonlítást a felületi örvények módszerével nyert eredményekkel tettük, és igen jó egyezést találtunk, amelyről a [9]-ben számoltunk be. Ugyancsak kiváló egyezést tapasztaltunk más szerzők számított eredményeivel az ellenállás-tényező falsúrlódásból származó részének tekintetében is (lásd [13]).

Az álló hengerre vonatkozó összehasonlítások mellett számítási eredményeink hossz-, illetve keresztirányú rezgést, valamint körmozgást végző hengerekre vonatkozóan más szerzők számítási eredményeivel is összehasonlítottuk (lásd [13]), amelyek közül csak egyet kívánunk itt bemutatni. A 6(a) és 6(b) ábrákon a tehetetlenségi erőt nem tartalmazó (lásd [8]) C_D ellenállás-tényező és a C_L felhajtóerő-tényező időbeli változásai láthatók keresztirányban rezgő henger esetén. A 6(a) ábra a szerző saját számítási eredményeit mutatja a durvább háló és időlépcső (301x177; $\Delta t = 0,0005$; lásd továbbá a 2. ábrát mindkét háló esetére), a 6(b) ábra pedig Lu és Dalton [26] ugyanerre az esetre vonatkozó eredményeit mutatja egy finomabb és egy durvább háló esetére. Látható, hogy a két számítás eredménye igen hasonló; a [26]-ban mindkét számítási háló esetén $\bar{C}_D = 1,25$, ill. $C_{Lrms} = 0,18$ értéket adnak meg, a jelen számítások pedig finom háló esetén $\bar{C}_D = 1,244$, ill. $C_{Lrms} = 0,185$, durva háló esetén pedig $\bar{C}_D = 1,243$, ill. $C_{Lrms} = 0,185$ értékeket szolgáltattak. A 6(a) és 6(b) ábra vízszintes tengelyein lévő értékek különbözősége az idő dimenziótlanításának különböző voltából származik.

Hasonlóan jó egyezést nyertünk akkor is, amikor eredményeinket Al-Mdallal és szerzőtársai [1] által a főáramlás irányában rezgőmozgást végző henger, ill. a körpályán keringő henger (lásd [19]) körüli áramlásra vonatkozó eredményeivel hasonlítottuk össze (lásd [13]). Miután meggyőződtünk róla, hogy a párhuzamos áramlásba helyezett álló, hossz- és keresztirányban rezgő, valamint a körpályán keringő henger körüli áramlásra jól működik a kifejlesztett számítási eljárás, rátértünk a párhuzamos áramlásba helyezett ellipszis pályán keringő henger körüli



(a) Baranyi (b) Lu és Dalton (1996)

6. ábra. A tehetetlenségi erőtől mentes felhajtóerő- és ellenállás-tényező időbeli változása (a) a szerző és (b) Lu és Dalton [26] számításai alapján (Re = 185, $A_x = 0$; $A_y = 0,2$, $f = 0,8\, rms\, St_0$, $\mathrm{St}_0 = 0,195$)

áramlás vizsgálatára. A gyakorlatban előfordul, hogy a henger nemcsak keresztvagy hosszirányban, hanem mindkét irányban rezeg. A vizsgálandó ellipszis pálya a henger hossz- és keresztirányú rezgésének szuperpozíciójaként állítható elő.

5. A henger ellipszis pályán történő mozgása

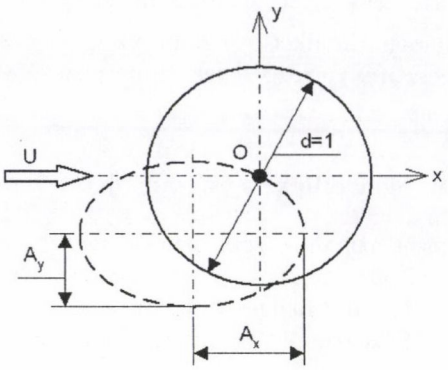
A 7. ábrán az ellipszis pályán mozgó körhengerre vonatkozó elrendezés látható. Az ellipszispályát két azonos frekvenciájú harmonikus rezgőmozgás eredőjeként kapjuk. Az egységnyi átmérőjű henger (minden hosszúságot a d hengerátmérővel dimenziótlanítunk) x_0, y_0 középpontjának mozgása a következő módon írható le:

$$x_0(t) = A_x \cos(2\pi f_x t) \qquad y_0(t) = A_y \sin(2\pi f_y t), \tag{30}$$

ahol az U/d-vel dimenziótlanított frekvenciák azonosak: $f_x = f_y = f$, az A_x, A_y az ellipszis dimenziótlan nagy- és kistengelye. A (30) egyenletet a t dimenziótlan idő szerint egyszer vagy kétszer deriválva megkapjuk a henger $\mathbf{v}_0$ sebességének és $\mathbf{a}_0$ gyorsulásának komponenseit. Természetesen ebben az esetben, mivel a henger forgómozgást nem végez, minden pontjának azonos a sebessége. Ha az A_x és A_y is nullától különböző, akkor a (30) egyenletből ellipszist kapunk (szaggatott vonal jelzi a 7. ábrán), amelynek az $e = A_y/A_x$ ellipticitása az A_y növelésével (rögzített A_x mellett) nő. Az $e = 0$ esetén a henger középpontja csak longitudinális rezgést végez, míg $e = 1$ esetén körpályán mozog. A (30) egyenlet óramutató járásával ellentétes hengermozgást eredményez; az y_0 előjelének megváltoztatásával óramutató járásával egyező bolygómozgást kapunk.

Minden egyes számítási sorozat esetében az Re Reynolds-számot, a rezgések A_x amplitúdóját valamint az f dimenziótlan frekvenciáját rögzített értéken tart-

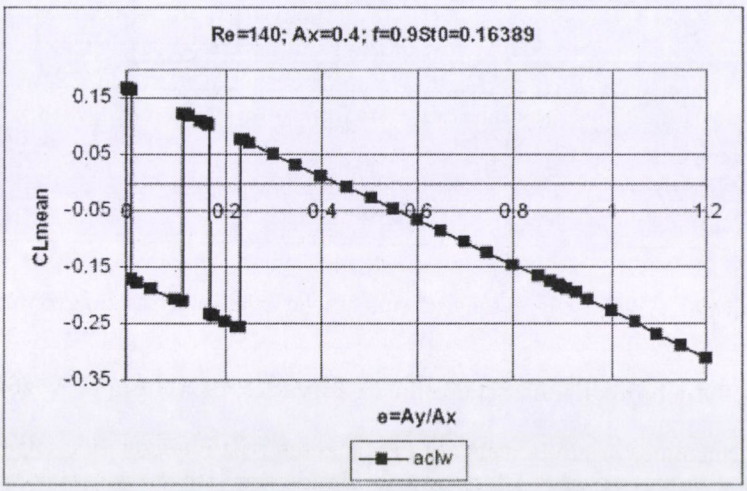
juk, és az A_y amplitúdót (s ezzel az e ellipticitást) változtatjuk. Az f értékét úgy választjuk meg, hogy az közel essen az adott Reynolds-számú, álló hengerről leváló örvények St_0 dimenziótlan örvényleválási frekvenciájához, és hogy a szinkronizálódás (az angol szakirodalomban lock-in-nak nevezett jelenség) a teljes vizsgált A_y, ill. e tartományban fennálljon. A lock-in állapot jellemzője, hogy az örvényleválás frekvenciája megegyezik (szinkronizálódik) a henger rezgésének frekvenciájával. A vizsgálatainkat a jelen tanulmányban $110 < \mathrm{Re} < 190$ Reynolds-szám tartományban végeztük. $\mathrm{Re} < 110$ esetén általában nem jelentkeztek azok az ugrások, amelyek tanulmányunk vizsgálatának tárgyai. Számítási tapasztalatunk azt mutatja, hogy egy bizonyos Reynolds-szám küszöb alatt nem alakul ki ez a jelenség. Mivel a [25] és [32] dolgozatokban bebizonyították, hogy a rezgés erősíti az áramlás kétdimenziós voltát, így ott a három-dimenziós jelenségek nagyobb Reynolds-számnál jelentkeznek, mint álló körhenger esetén. Ezáltal biztosított, hogy az adott Re tartományban az áramlás kétdimenziós, és így a számítási eljárásunk ezekre az esetekre megbízhatóan használható. Ezt támasztja alá a [11] dolgozatunk is, amelyben az $\mathrm{Re} \leq 300$ Reynolds-szám tartományban végzett vizsgálataink során az itt bemutatottakhoz hasonló eredményekre jutottunk.



7. ábra. Henger ellipszis pályán történő keringése

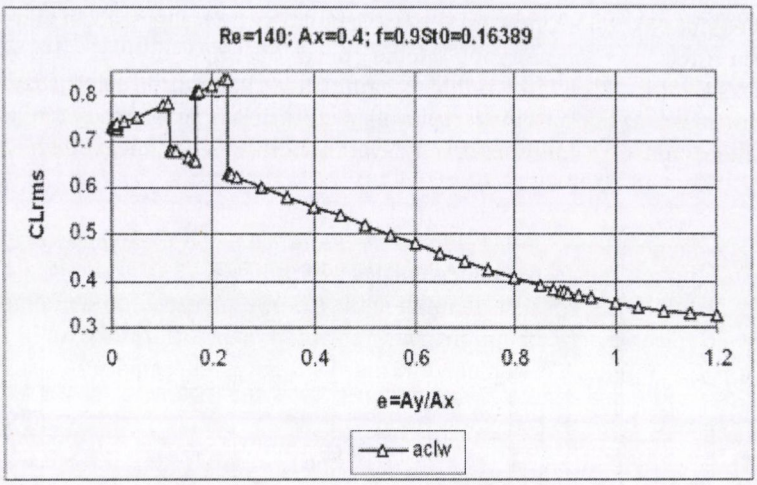
Érdekes jelenséget tapasztaltunk, amikor homogén párhuzamos áramlásba helyezett ellipszis pályán mozgó körhenger esetén a C_L felhajtóerő-tényező, a C_D ellenállás-tényező és a C_{pb} hátsó nyomástényező (base pressure coefficient) időátlagát és az oszcilláló jel amplitúdójára jellemző effektív középértéket (rms érték) vizsgáltuk rögzített Re, A_x és f értékek esetén. A jellemzőket az e ellipticitás függvényében felrajzolva, azok mindegyikében ugyanazon az e értékeknél hirtelen, ugrásszerű változást tapasztaltunk, lásd [6], [7], [10] és [13]. Egy tipikus példát mutat erre az óramutató járásával ellentétes irányban (aclw) keringő henger esetén a 8. ábra, ahol $\mathrm{Re} = 140$, $A_x = 0,4$, $f = 0,9\mathrm{St}_0$ ($\mathrm{St}_0 = 0,1821$) paraméterek mellett a C_L időátlagát ábrázoljuk az e ellipticitás függvényében. Az ábrán három ugrás látható. Mind az alsó, mind a felső görbe mentén – nevezzük őket a továbbiakban

*állapotgörbé*knek – a függvény értéke közel lineárisan csökken e növekedésével, az iránytangensük is közel azonos, így a két görbe közti távolság is közel állandó. Természetesen a felső görbén nagyobb a felhajtóerő időátlaga, mint az alsó görbén. A jelen szerző a [6] és [7] dolgozataiban megmutatta, hogy a C_L felhajtóerő-tényező időbeli változása az ugrás előtti és utáni e ellipticitás értékek esetén jelentősen különbözik egymástól, amely az örvényleválás szerkezetének megváltozására utal.



8. ábra. A felhajtóerő-tényező időátlaga az ellipticitás függvényében

Vizsgálataink azt mutatják, hogy amikor a fenti három mennyiség rms értékét, ill. a C_D és a C_{pb} mennyiségek időátlagát ábrázoljuk az e függvényében, akkor mind az öt esetben a 8. ábrán bemutatott esettől különböző, de egymáshoz abban hasonlító görbéket kapunk, hogy a két állapotgörbe az $e = 0$ értéknél metszi egymást és e értékének növekedésével a két állapotgörbe egyre jobban eltávolodik egymástól. Tehát, amennyiben a henger csak longitudinális rezgést végez ($e = 0$), akkor e elemi növelésével, a két állapotgörbe közti különbség is elemi marad, szemben a 8. ábrán bemutatottaktól. Természetesen $e > 0$ esetben elemi kis e változáshoz a függő változó véges nagyságú megváltozása tartozhat. Mivel az említett 5 jellemző változása egymáshoz hasonló, így közülük itt csak egy tipikus – a 9. ábrán látható – esetet, a $C_{Lrms}(e)$ függvényt mutatjuk be, amely ugyanazon paraméterértékekhez tartozik, mint a 8. ábra. Ez az eredmény is azt az elgondolást támogatja, hogy a vizsgált e tartományban két fajta örvényleválás fordulhat elő, és ezért a $C_L(t)$, $C_D(t)$ és $C_{pb}(t)$ függvények mindegyikének két megoldása lehet. Megjegyezzük, hogy mind a 6 említett esetben azonos számú ugrást találtunk, és az ugrások helye is azonos volt.



9. ábra. A felhajtóerő-tényező rms értéke az ellipticitás függvényében

6. Mechanikai energiaátadás a folyadék és a henger között

A párhuzamos áramlás irányára merőlegesen rezgő henger és a folyadék közti mechanikai energiaátadást Blackburn és Henderson [16] vizsgálta először. Az ő elméletüket terjesztjük itt ki a henger 2D mozgására. Az ellipszis pályán keringő henger esetén a folyadék és henger között longitudinális és transzverzális irányban is van mechanikai energiaátadás. Az E fajlagos *mechanikai energiaátadási tényező* értékét akkor vizsgáljuk, amikor az áramlás már periodikus. Így a dimenziótlan $x_0(t)$, $y_0(t)$ hengerelmozdulás koordinátáiból és a hengerre jellemző $C_L(t)$ és $C_D(t)$ erőtényezőkből a $(y_0(t), C_L(t))$ és $(x_0(t), C_D(t))$ határciklusok képezhetők. Az energiacserét akkor tekintjük pozitívnak, ha a hengeren történik a munkavégzés, azaz ha a folyadék energiát ad át a hengernek, és negatívnak, ha a henger ad át energiát a folyadéknak.

Terjesszük most ki most a [16] egy $\tilde{T}$ örvényleválási periódusra vonatkoztatott E mechanikai energiaátadási tényezőjének definícióját ellipszis pályán mozgó henger egységnyi hosszúságú szelvényére! Legyen

$$E = \frac{2}{\rho U^2 d^2} \int_0^{\tilde{T}} \mathbf{F} \cdot \mathbf{v}_0 U \, \mathrm{d}\tilde{t} = \frac{2}{\rho U^2 d^2} \int_0^{\tilde{T}} (F_D v_{0x} + F_L v_{0y}) \, \mathrm{d}\tilde{t} = \\ = \int_0^T (C_D \dot{x}_0 + C_L \dot{y}_0) \, \mathrm{d}t = E_2 + E_1, \tag{31}$$

tehát

$$E_1 = \int_0^T C_L(t) \, \dot{y}_0(t) \, \mathrm{d}t, \qquad E_2 = \int_0^T C_D(t) \, \dot{x}_0(t) \, \mathrm{d}t. \tag{32}$$

Megjegyezzük, hogy a (31) egyenletben az 1. és 2. egyenlőségjel utáni mennyiségek mind dimenziós fizikai mennyiségek, míg a 3. és 4. egyenlőségjel után már csak dimenziótlan mennyiségek szerepelnek. Az egyenletben ρ a folyadék sűrűsége, U a párhuzamos áramlás sebessége, d a henger átmérője, $\mathbf{F} = F_D\mathbf{i} + F_L\mathbf{j}$ az egységnyi hosszúságú hengerre ható erővektor, amelynek komponensei a párhuzamos áramlás irányában ható F_D ellenállás és az arra merőleges F_L felhajtóerő, $\mathbf{i}, \mathbf{j}$ az x, y irányba mutató egységvektorok, $\mathbf{v}_0 = v_{0x}\mathbf{i} + v_{0y}\mathbf{j}$ a henger középpontjának sebességvektora, $\tilde{t}$ [s] az idő, $\tilde{T}$ [s] az örvényleválás periódusideje, míg a nekik megfelelő t és T mennyiségek a d/U mennyiséggel vannak dimenziótlanítva. A (31) egyenletben szereplő C_L és C_D a Jelölésjegyzékben definiált felhajtóerő-tényező, ill. ellenállás-tényező, $\dot{x}_0$ és $\dot{y}_0$ pedig a henger középpontjának U-val dimenziótlanított x, ill. y irányú sebességkomponense. Az egyenletből látható, hogy az energiaátadás az (x, y) irányokban vett energiacserék összegeként állítható elő: $E_1 + E_2$. Az E mechanikai energiaátadási tényező (31) definíciója abban a határesetben, amikor a henger csak keresztirányú vagy transzverzális (y) rezgést végez, éppen a [16] által bevezetett energiaátadási tényező összefüggését szolgáltatja: $E = E_1$. Csak longitudinális (x irányú) rezgést végző henger esetén $E = E_2$ adódik. Az E_1 és E_2 értékeit a (32) egyenletek alapján számítjuk a t dimenziótlan időben egyenközű osztásban, diszkrét pontokban adott függvények numerikus integrálása segítségével.

A Stokes-tétel síkbeli esetre vonatkozó speciális alakja, a Green-tétel (lásd például [21] és [30] könyveket) felhasználásával az E_1 és E_2 energiaátadási tényezőknek geometriai jelentést is tulajdoníthatunk. A Green-tétel alkalmazásával (amely többek között síkidomok területének vonalintegrállal történő kiszámítására is használható) az E_1 átalakítható:

$$E_1 = \int_0^T C_L(t)\,\dot{y}_0(t)\,\mathrm{d}t = \oint C_L \mathrm{d}y_0 = -\oint y_0 \mathrm{d}C_L = \\ = \frac{1}{2}\left(\oint C_L \mathrm{d}y_0 - \oint y_0 \mathrm{d}C_L\right), \tag{33}$$

ahol a vonalintegrálokat az óramutató járásával egyező irányban kell elvégezni a (y_0, C_L) határciklust jelentő zárt görbe mentén. Ehhez hasonlóan az x irányú energiaátadási tényező összefüggése is átalakítható:

$$E_2 = \int_0^T C_D(t)\,\dot{x}_0(t)\,\mathrm{d}t = \oint C_D \mathrm{d}x_0 = -\oint x_0 \mathrm{d}C_D = \\ = \frac{1}{2}\left(\oint C_D \mathrm{d}x_0 - \oint x_0 \mathrm{d}C_D\right). \tag{34}$$

A vonalintegrálokat itt is az óramutató járásával egyező irányban kell elvégezni az (x_0, C_D) határciklust jelentő zárt görbe mentén.

Az E_1 és E_2 mennyiségek geometriai jelentése az (y_0, C_L), ill. (x_0, C_D) határciklusok által határolt előjelhelyes terület. Bármely mennyiség akkor pozitív, ha a hozzá tartozó határciklus görbéje az óramutató járásával megegyező irányítású.

Bár az E_1 és E_2 értékeit a (32) egyenletek alapján célszerű meghatározni, a (33) és (34) egyenletek szemléletes geometriai jelentést adnak e mennyiségeknek. A (31) egyenlet alapján a körhenger és a folyadék közti teljes energiaátadási tényező:

$$E = E_2 + E_1.$$

7. Számítási eredmények és diszkusszió

A vizsgálatok elvégzéséhez igen nagy számú esetet vettünk figyelembe. A henger dimenziótlan $f = f_x = f_y$ rezgési frekvenciáját a $0,7\mathrm{St}_0 < f < 1,1\mathrm{St}_0$ között $\Delta f = 0,05\mathrm{St}_0$ lépcsőzéssel változtattuk, ahol St_0 az adott Reynolds-számhoz tartozó, álló henger dimenziótlan örvényleválási frekvenciája. Még egy adott frekvenciához tartozóan is sok számítást végeztünk. Például az $f = 0,9\mathrm{St}_0$ frekvenciánál 5 különböző számítási sorozatot értékeltünk ki: Re = 120-nál és Re = 140-nél $A_x = 0,4$; Re = 160-nál $A_x = 0,2$ és $A_x = 0,3$; Re = 180-nál $A_x = 0,3$. Az említett ötből itt most csak egy esetet mutatunk be, mivel a többi is hasonló jellegű eredményeket adott.

7.1. Energiaátadásra vonatkozó eredmények

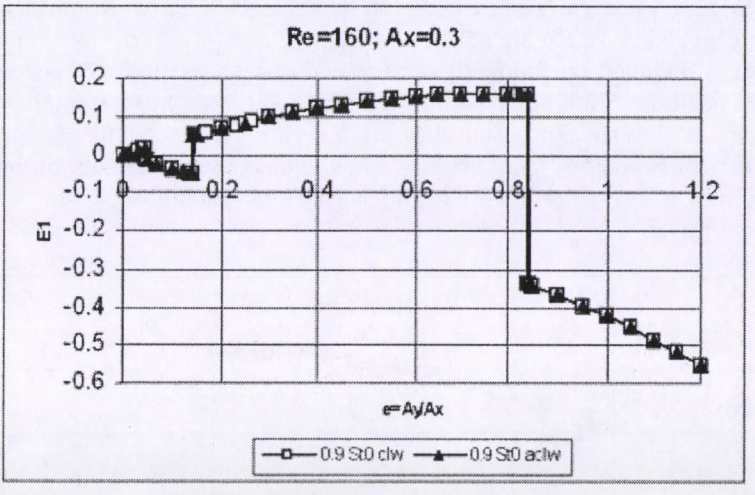
A bemutatott számítási esetet az

$$\mathrm{Re} = 160, \quad A_x = 0,3, \quad \text{és} \quad f = 0,9\mathrm{St}_0 = 0,16938$$

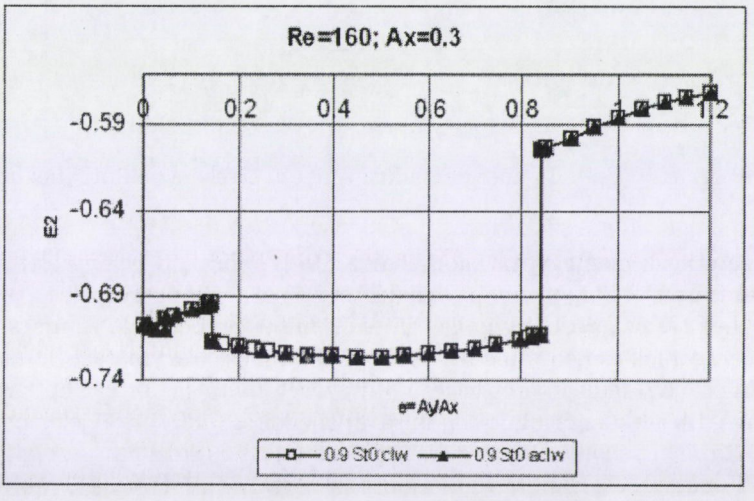
paraméterek jellemzik. A számításokat mind óramutató járásával egyező (*clw*), mind azzal ellentétes (*aclw*) irányban keringő henger esetére elvégeztük. Az e ellipticitás értéke egy számításon belül rögzített. Az e értékeit a 0 (tisztán longitudinális rezgés) értéktől az 1,2 értékig (körpályán túl) úgy választottuk, hogy viszonylag egyenletesen és sűrűn lefedje a vizsgált tartományt, és így minden ugrást ki tudjunk mutatni. Amikor egy ugrást találtunk, akkor annak mindkét oldalán számos újabb számítást végeztünk az ugrás helyének lokalizálása céljából. A kezdeti feltételt mindkét irányú keringés esetén azonosra választottuk: $x_0\,(t = 0) = A_x$, $y_0\,(t = 0) = 0$. A nagy számú számítási eredményből itt csak egy jellemző példát mutatunk be.

A 10–12. ábrák a fajlagos mechanikai energiaátadási tényezők (E_1, E_2, E) változását mutatják az e ellipticitás függvényében a fent említett paraméterek mellett mind az óramutató járásával azonos, mind azzal ellentétes irányban keringő henger esetén. A 10. ábra a henger transzverzális mozgásösszetevőjéből származó, folyadék és test közötti E_1 energiaátadási tényezőt mutatja. Az üres négyzet jelek az óramutató járásával egyező (az ábrán „*clw*"-vel jelölve), a tömör háromszögek pedig az azzal ellentétes (az ábrán „*aclw*"-vel jelölve) irányú keringéshez tartozó számítási eredményeket mutatják. Mindkét keringési irányhoz egy pár burkológörbe vagy állapotgörbe tartozik, amelyek $e = 0$ értéknél metszik egymást, és ez a két pár

burkológörbe vonalvastagságon belül megegyezik egymással. Az ábráról az is látható, hogy az ugrások helye is azonos a két keringési irány esetén. Vegyük észre, hogy a felső állapotgörbén $E_1 > 0$, ami azt jelenti, hogy a folyadék energiát ad át a hengernek. Ugyanakkor az alsó állapotgörbén $E_1 < 0$; ilyenkor az energiaátadás fordított irányú és ez a henger mozgását fékezni igyekszik.



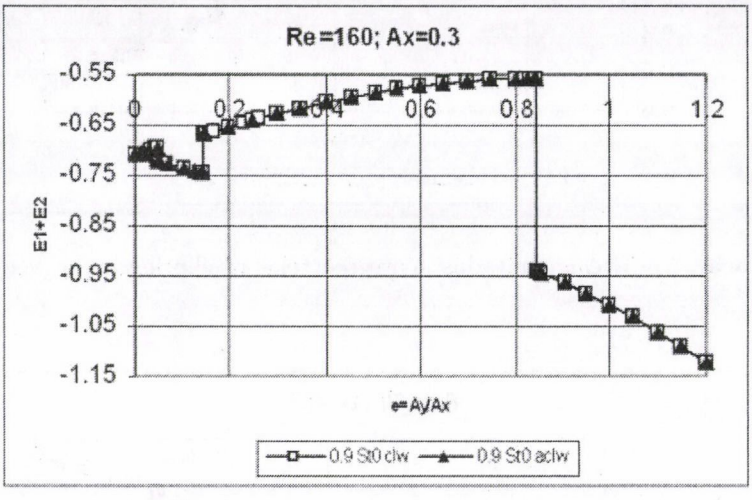
10. ábra. Az E_1 energiaátadási tényező értéke az ellipticitás függvényében



11. ábra. Az E_2 energiaátadási tényező értéke az ellipticitás függvényében

A 11. ábra a henger longitudinális irányú mozgásösszetevőjéből származó, folyadék és test közötti E_2 energiaátadási tényezőt mutatja az e függvényében. A görbék részben hasonlóak a 10. ábrán bemutatottakhoz (két pár egybeeső burkológörbe, azonos helyen lévő ugrások), ugyanakkor megfigyelhető, hogy az E_2 értékek mindkét burkológörbén negatívak. Ez azt jelenti, hogy a folyadék energiát nyer a hengertől, és fékezni igyekszik annak mozgását.

Az E_1 és E_2 energiaátadási tényezők E összegét a 12. ábra mutatja. A két burkológörbe alakja az E_1 alakjához hasonló, de az E értékek – az E_2 értékekhez hasonlóan – mindkét burkológörbén negatívak. Így egy keringő mozgást végző henger és az áramló folyadék között a mechanikai energiacsere mindig negatív, azaz a henger ad át energiát a folyadéknak a keringés irányától függetlenül a teljes vizsgált e tartományban. Egy örvényleválási ciklus esetén tehát a mozgó henger végez munkát a folyadékon, és a folyadék egyfajta ellenállást gyakorol a henger mozgásával szemben.

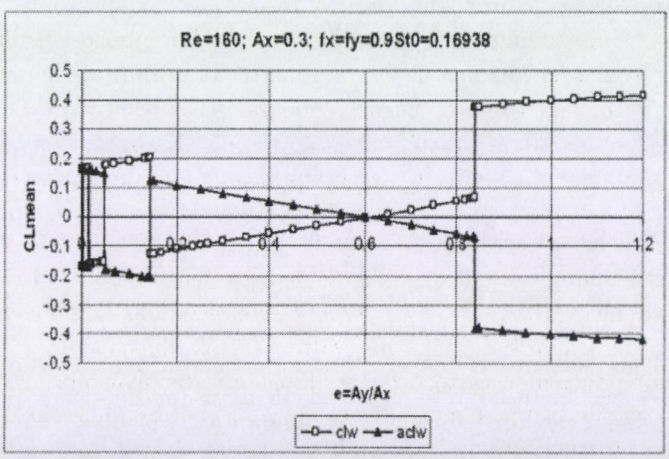


12. ábra. Az $E = E_1 + E_2$ energiaátadási tényező értéke az ellipticitás függvényében

Itt szeretnénk megjegyezni, hogy bár a 10–12. ábrákon bemutatott erőtényezők (és az összes többi, 9. ábrán bemutatott típusú állapotgörbék is) két keringési irányhoz tartozó görbéi szimmetria okok miatt egybeesnek, ez nincs így azon típusú határgörbék esetén, amelyek közel párhuzamosak egymással (lásd 9. ábra). A 13. ábra a C_L felhajtóerő-tényező változását mutatja az e ellipticitás függvényében. A ▲ jelek az óramutató járásával ellentétes (*aclw*), a □ jelek pedig azzal megegyező irányítású (*clw*) bolygómozgás esetére vonatkoznak. Az *aclw* esethez tartozó állapotgörbék iránytangense most is negatív (mint azt a 8. ábrán is láttuk), az óramutató járásához (*clw*) tartozóan viszont pozitív iránytangenst kapunk. Vegyük észre, hogy mindkét mozgáshoz egy-egy pár állapotgörbe tartozik, amelyek

most is páronként közel párhuzamosak egymással. Az ábrát közelebbről megvizsgálva feltűnik, hogy a két görbe a vízszintes tengelyre vett tükörképe egymásnak. Az, hogy a két keringési irányhoz tartozó felhajtóerőnek különböznie kell egymástól, könnyen belátható. A *clw* esetben, amikor a henger a pályája legmagasabb $(0, Ay)$, ill. legalacsonyabb $(0, -Ay)$ pontjában van, akkor a henger x irányú maximális sebessége ($u_{0\max} = 2\pi f Ax$; lásd a (30) egyenlet t szerinti deriváltját) hozzáadódik az U megfúvási sebességhez, ill. kivonódik abból. Az ellenkező keringési irány (*aclw*) esetén ez éppen fordítva van. A különböző sebességek miatt más a test mentén kialakuló nyomás és a nyírófeszültség, amely a mozgás szimmetriáját is tekintve testre ható függőleges erők különbözőségéhez vezet. A 10–13. ábrákról látható, hogy a forgásirány megváltoztatása nem változtatja meg az ugrások helyét és számát.

Szemben a hengermozgás irányításának megváltoztatásával, a henger mozgására vonatkozó kezdeti feltételben (amely például egy szöggel jellemezhető; jelöljük Θ-val) történő változtatás megváltoztathatja azokat az e ellipticitás értékeket, amelyeknél az örvénystruktúra ugrásszerűen változik meg. Ennek a vizsgálatával is foglalkoztunk a [13] dolgozatban, de erre itt most nem kívánunk kitérni. Az e ellipticitás (és egyéb paraméterek) kis megváltozása esetén bizonyos esetekben fellépő ugrásszerű megváltozás egy megközelítése az, hogy a vizsgált nemlineáris problémához valószínűleg két periodikus attraktor (*attractor*) tartozik egy-egy vonzási tartománnyal (*basin of attraction*). Az e ellipticitási paraméter értékének megváltoztatásával az attraktorokat elválasztó határ (*basin boundary*) másik oldalára kerülhet a rendszer, a megoldás a másik attraktorhoz vonzódik és ezáltal ugrásszerűen megváltozhat az örvényszerkezet (lásd a [18], [27] és [29] könyveket). A gondolatot kiterjesztve azt mondhatjuk, hogy a párhuzamos áramlásba helyezett ellip-



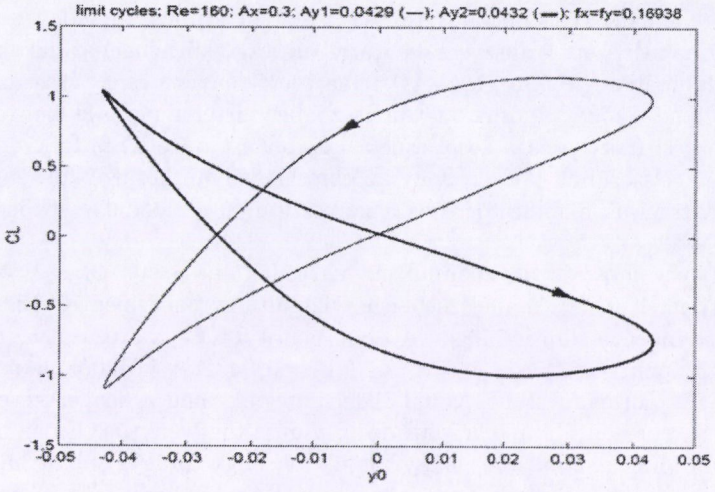
13. ábra. A keringés irányának hatása a felhajtóerő-tényező időátlagára

szis pályán mozgó henger körüli áramlást jellemző 5 elemű (Re, A_x, e, f és Θ) paraméter-rendszer egy vagy több elemének megváltoztatása megváltoztathatja azt, hogy a megoldás a nemlineáris rendszer melyik attraktorához vonzódik, (lásd [12]).

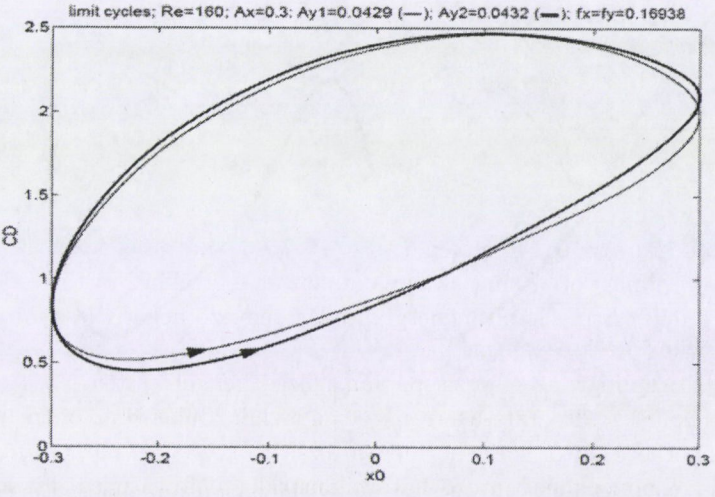
A következőkben az ugrásokat kívánjuk vizsgálatunk tárgyává tenni. A jelen szerző a [6]-ban a C_L felhajtóerő-tényező időbeli változását vizsgálva azt találta, hogy közvetlenül az ugrás előtti és utáni e értékekhez tartozó $C_L(t)$ függvények jelentősen különböztek egymástól, amelyek a jelek különböző időátlagát és *rms* értékét eredményezték. Itt a henger óramutató járásával egyező irányban történő keringése esetén az ugrás előtti és utáni határciklusokat vizsgáljuk a jelen alfejezet elején definiált paraméterek esetén. A vizsgált ugrás helye $e \approx 0,1435$ (lásd a 10. ábrát). A 14. ábra két (y_0, C_L) határciklust mutat, amely egy ellipszis pályán mozgó henger transzverzális mozgásirányára vonatkozik. A vékony vonal az ugrás előtti e értékhez ($e = 0,143$; $A_y = 0,0429$), a vastag vonal pedig az ugrás utáni e értékhez ($e = 0,144$; $A_y = 0,0432$) tartozó határciklust mutatja. Bár a két e érték majdnem azonos, a két határciklus görbe jelentősen különbözik egymástól. A görbék közel egymás tükörképei, és irányításuk is ellentétes. Ez azt jelenti, hogy a két esetben az energiaátadás előjele különböző: $e = 0,143$ esetén negatív ($E_1 = -0,0521$) és az $e = 0,144$ esetén pozitív ($E_1 = 0,0491$). Ez az eredmény hasonló ahhoz, amit [16] és [17] a párhuzamos áramlásba helyezett transzverzális irányban változó frekvenciával rezgő henger esetében tapasztalt. Ezekben a tanulmányokban azt találták, hogy egy kritikus frekvencia értéken áthaladva a határciklus görbék irányítása ellentétessé válik, amely különböző előjelű energiacserét jelent.

A 15. ábra két (x_0, C_D) határciklust mutat, amely az ellipszis pályán mozgó henger longitudinális mozgására vonatkozik. Itt is ugyanahhoz a két e értékhez tartozó határciklusokat ábrázoljuk, mint a 14. ábrán. Az ábráról látható, hogy – szemben a 14. ábrán bemutatott görbékkel – a két különböző e értékhez tartozó határciklus görbéi közel azonosak, és mindkettő az óramutató járásával ellentétes irányítású. Ez az irányítás negatív energiacserét jelent: $e = 0,143$ esetén $E_2 = -0,6947$ és $e = 0,144$ esetén $E_2 = -0,7663$. Látható, hogy az E_2 abszolút értéke sokkal nagyobb, mint az ugyanahhoz az e értékhez tartozó E_1 abszolút értéke, így a teljes E mechanikai energiaátadási tényező negatív lesz mindkét vizsgált e ellipticitás érték esetében.

Azt találtuk, hogy az ellipszis pályán mozgó henger esetén, az e ellipticitás egy kis mértékű megváltoztatása a transzverzális elmozdulás-komponenshez tartozó határciklus görbe jelentős mértékű megváltozását okozhatja, míg a longitudinális elmozdulás-komponenshez tartozó határciklust alig befolyásolja. Ez azt jelenti, hogy a transzverzális irányú elmozdulás-erő (y_0, C_L) határciklus sokkal érzékenyebb arra a jelenségre, amely a korábban említett ugrást okozza, mint a longitudinális irányú elmozdulás-erő (x_0, C_D) határciklus. Ez azt sugallja, hogy a felhajtóerő és ellenállás más jellemzőkre, így például a fázisszögre is különböző hatást gyakorol.



14. ábra. Az (y_0, C_L) határciklus görbék az ugrás előtti és utáni ellipticitás érték esetén

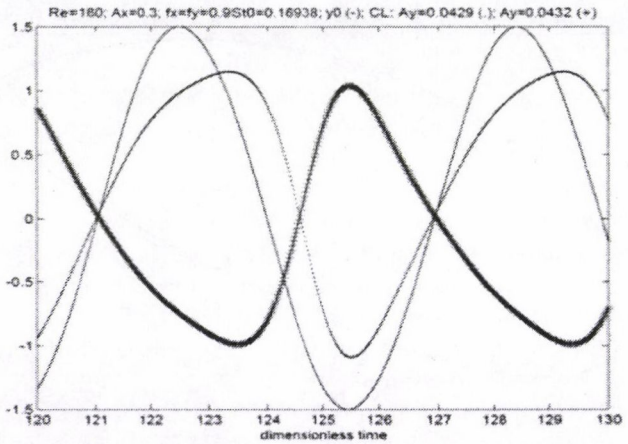


15. ábra. Az (x_0, C_D) határciklus görbék az ugrás előtti és utáni ellipticitás értékek esetén

7.2. Fázisszöggel kapcsolatos eredmények

Több tanulmány, így például [16], [17] és [26] kimutatta, hogy a párhuzamos áramlásba helyezett transzverzális irányban rezgő körhenger szinkronizálódott áramlási állapotában (lock-in) a $C_L(t)$ felhajtóerő-tényező és $y_0(t)$ transzverzális henger-elmozdulás között mérhető Φ_L fázisszögben hirtelen ugrás léphet föl, amikor a henger rezgési frekvenciája közel esik a korábban említett St_0 frekvenciájához. Ezek alapján célszerűnek látszik, hogy az ellipszis pályán keringő körhenger esetében is megvizsgáljuk a felhajtóerő és a keresztirányú elmozdulás között mérhető Φ_L fázisszöget.

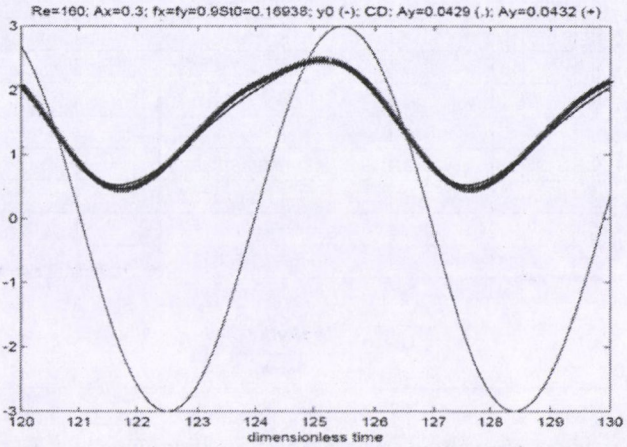
Nézzük meg most is ugyanannak az ugrásnak ($e_1 = 0,143$, $e_2 = 0,144$) a környezetét, amelyet a 7.1. alfejezetben is vizsgáltunk (lásd még 10–13. ábrákat). A 16. ábra a dimenziótlan idő függvényében mutatja a C_L változását az $e = 0,143$ ($A_y = 0,0429$; pontvonal) és az $e = 0,144$ ($A_y = 0,0432$; + jelekből alkotott vonal) értékeknél. Az ábrán folytonos vonallal feltüntettük még a henger keresztirányú elmozdulását reprezentáló 1,5 amplitúdójú szinusz függvényt is. Ez az amplitúdó a valóságban jóval kisebb; azért nagyítottuk fel, hogy az így jobban használható legyen a Φ_L fázisszög vizsgálatakor. Az ábrán látható, hogy míg az ugrás előtti ellipticitás értékhez ($e = 0,143$) tartozó C_L gyakorlatilag fázisban van a hengerelmozdulással, addig az ugrás utáni ($e = 0,144$) értékhez tartozó C_L közelítőleg ellenfázisban van a hengerelmozdulással. Így tehát a hengerelmozdulás és felhajtóerő-tényező közötti Φ_L fázisszög mintegy 180°-os változást szenvedett, miközben az ugráson áthaladva az e érték minimálisan változott.



16. ábra. A C_L időbeli változása az ugrás előtti és utáni ellipticitás értékekre

A 17. ábra tanúsága szerint a henger longitudinális $x_0(t)$ elmozdulása és a $C_D(t)$ ellenállás-tényező között mérhető fázisszög elhanyagolható mértékben változik az ugrás hatására. Most is folytonos vonal jelzi a henger $x_0(t)$ longitudi-

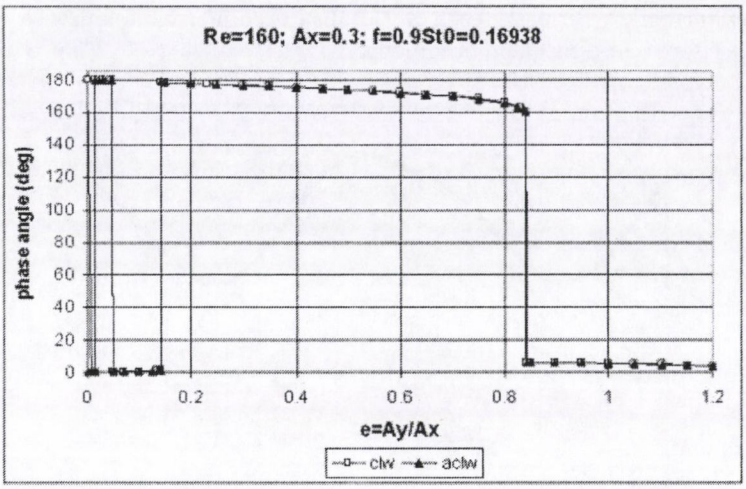
nális elmozdulását. A két másik – alig megkülönböztethető – görbe az ugrás előtti ($e = 0,143$; pontvonal) és utáni ($e = 0,144$; + jelekből alkotott vastag vonal) e értékekhez tartozik. Természetesen a 15. ábra eredményeit tekintve a két görbe közötti kis eltérés nem meglepő eredmény.



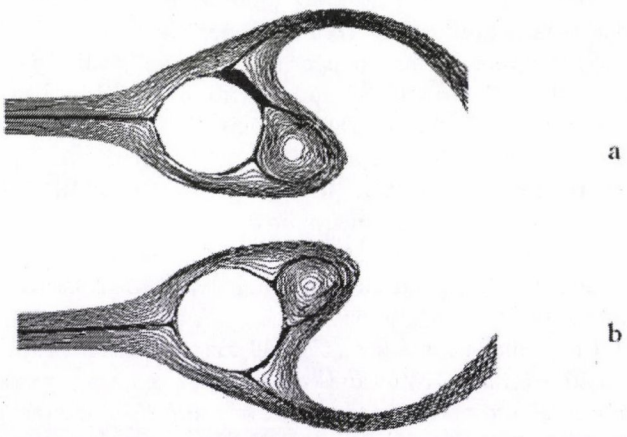
17. ábra. A C_D időbeli változása az ugrás előtti és utáni ellipticitás értékekre

Ezek az eredmények azt sugallják, hogy a továbbiakban a henger keresztirányú elmozdulása és a felhajtóerő-tényező között mérhető Φ_L fázisszögre érdemes koncentrálni. Ezért a 7.1. és jelen alfejezetekben részletesen bemutatott esetre vonatkozóan meghatároztuk a Φ_L változását az e ellipticitás függvényében (18. ábra) mind az óramutató járásával egyező (*clw* üres négyzet) és azzal ellentétes (*aclw* töltött háromszög) irányban keringő henger esetén. Ezen az ábrán is ugyanazoknál az e ellipticitás értékeknél vannak az ugrások, mint a 10–12. ábrákon. Az ugrásokon keresztül haladva gyakorlatilag 180°-os fázisszög változás látható a 18. ábrán. Megnyugtató, hogy a két keringési irányhoz tartozó, egymástól függetlenül nyert, fázisszög értékek teljesen egybeesnek (lásd [13]). A 10. és 18. ábrákra tekintve látható, hogy az E_1 tényező előjele meghatározza a Φ_L fázisszög értékét: $E_1 > 0$ esetén $\Phi_L \approx 180°$, míg $E_1 < 0$ esetén $\Phi_L \approx 0°$. Ezek az eredmények összhangban vannak [16], [17] és [26] párhuzamos áramlásba helyezett keresztirányban rezgetett henger fázisszögére vonatkozó eredményeivel.

A [16] és [17] a párhuzamos áramlásba helyezett, transzverzális irányban rezgetett henger körüli áramlásra vonatkozó egy hipotézise az, hogy áramlás struktúrájának ugrásokhoz vezető megváltozását két különböző örvénytermelő mechanizmus közötti egyensúly megváltozása okozza. Ezt a hipotézist a vizsgálataik során nyert eredményeikkel támasztják alá, és úgy tűnik, hogy a párhuzamos áramlásba helyezett keringő mozgást végző henger esetében ugyanez lehet a korábban említett, ugrásokat előidéző jelenség mögött. A 19a és 19b ábrák a numerikus eredményekből származtatott, a hengerhez kötött rendszerben értelmezett áramvonalakat mutat-



18. ábra. A Φ_L fázisszög változása az ellipticitás függvényében óramutató járásával egyező (clw) és azzal ellentétes (aclw) irányban keringő henger esetén



19. ábra. Azonos időpontbeli áramképek. a: $e = 0,143$; b: $e = 0,144$

ják a korábban vizsgált ugrás előtti ($e = 0,143$) és ugrás utáni ($e = 0,144$) ellipticitás értékek esetén. Mindkét áramkép ugyanahhoz az időponthoz tartozik (a henger egy felső transzverzális holtpontjához tartozó idő). Ez a két ábra is jól illusztrálja, hogy az e ellipticitás kis mértékű megváltozása az örvényleválás mechanizmusának milyen markáns megváltozását okozhatja. Az ábrákon feltüntetett esetben a két áramkép közel egymás tükörképe, s ez összhangban van a korábban vizsgált Φ_L fázisszög 180°-os ugrásszerű megváltozásával. A párhuzamos áramlásba helyezett keresztirányban rezgetett henger esetében az áramképre vonatkozóan hasonló eredményre jutott [17] is.

8. Összefoglalás

Ebben a dolgozatban kiterjesztettük Blackburn–Henderson [16], a párhuzamos áramlásba helyezett transzverzális irányban rezgőmozgást végző körhenger és folyadék között értelmezett E mechanikai energiaátadási tényezőjét az elliptikus pályán mozgó körhenger esetére. Ebben az esetben az E az E_1 és az E_2 értékekből tevődik össze, amelyek a transzverzális (E_1), ill. a longitudinális (E_2) mozgás- és erőkomponensekből származnak. A dolgozatban megvizsgáltuk, hogyan függnek az E_1, E_2 és $E = E_1 + E_2$ mennyiségek az e ellipticitástól. Két burkológörbét, ill. állapotgörbét találtunk, amelyek között a megoldások ugrásszerűen változnak. Az E_1 értékek pozitív és negatív értékeket is felvesznek, míg az E_2 mindig negatív. Ugyancsak mindig negatív az $E = E_1 + E_2$ eredő mechanikai energiaátadási tényező is. Az (y_0, C_L) és (x_0, C_D) határciklusokat megvizsgáltuk egy-egy, közvetlenül az ugrás előtti és utáni, egymástól alig különböző e értéknél. Az (y_0, C_L) határciklusról azt találtuk, hogy a két e értékhez egymástól teljesen különböző két görbe tartozik, és még a görbe irányítása is megváltozik, amely az E_1 tényező előjelét is megváltoztatja. Ugyanakkor az (x_0, C_D) határciklus görbéi alig változnak e kis mértékű megváltoztatásának hatására. Az itt említett e értékekre ugyanabban az időpontban felrajzoltuk a számított áramképeket is. Az áramképek különbözősége jól illusztrálja azt a tényt, hogy az e ellipticitás kis mértékű megváltozása az örvényleválás mechanizmusának milyen markáns megváltozását okozhatja.

A $C_L(t)$ felhajtóerő-tényező és az $y_0(t)$ transzverzális irányú hengerelmozdulás közötti Φ_L fázisszög vizsgálata azt mutatta, hogy minden egyes ugráson áthaladva a fázisszög mintegy 180°-os változást szenved. Azt találtuk, hogy az E_1 tényező előjele meghatározza a Φ_L fázisszög értékét: $E_1 > 0$ esetén $\Phi_L \approx 180°$, míg $E_1 < 0$ esetén $\Phi_L \approx 0°$. Ugyanakkor nincs fázisszög eltolódás a $C_D(t)$ és $x_0(t)$ jelek között az ugráshoz tartozó e értéken történő áthaladáskor.

Ezek az eredmények annak a jelenségnek a létezését támasztják alá, amelyek a kis Reynolds-számú, ellipszis pályán mozgó körhenger esetén bizonyos jellemzők ugrásszerű megváltozását idézik elő. Valószínűleg a vizsgált nemlineáris problémához két periodikus attraktor tartozik egy-egy vonzási tartománnyal. Az e ellipticitási paraméter értékének megváltoztatásával az attraktorokat elválasztó határ

másik oldalára kerülhet a rendszer, a megoldás a másik attraktorhoz vonzódik és ezáltal ugrásszerűen megváltozhat az örvényszerkezet. Az eredmények a felhajtóerőtényező időtől való függését, a határciklusokat, a mechanikai energiaátadási tényezőt, fázisszöget és áramképeket is magukba foglalnak.

9. Köszönetnyilvánítás

A szerző köszönetet mond a jelen kutatáshoz kapott OTKA (projekt szám: T 042961) támogatásért és Ujvárosi Sándor úrnak a kutatáshoz nyújtott segítségéért.

Hivatkozások

[1] Al-Mdallal, Q.M., Lawrence, K.P. and Kocabiyik, S.: *Forced streamwise oscillations of a circular cylinder: Locked-on modes and resulting fluid forces*, Journal of Fluids and Structures **23**(5) (2007) 681–701.

[2] Anderson, J.D., Jr.: *Computational Fluid Dynamics* (McGraw-Hill, New York, 1995).

[3] Baranyi, L. and Shirakashi, M.: *Numerical solution for laminar unsteady flow about fixed and oscillating cylinders*, Journal of Computer Assisted Mechanics and Engineering Sciences **6** (1999) 263–277.

[4] Baranyi, L.: *Numerical analysis of unsteady viscous flow around heat exchanger elements*, Acta Mechanica Slovaca 2/2002 (2002) 485–490.

[5] Baranyi, L.: *Computation of unsteady momentum and heat transfer from a fixed circular cylinder in laminar flow*, Journal of Computational and Applied Mechanics **4**(1) (2003) 13–25.

[6] Baranyi, L.: *Numerical simulation of flow past a cylinder in orbital motion*, Journal of Computational and Applied Mechanics **5**(2) (2004) 209–222.

[7] Baranyi, L.: *Sudden jumps in time-mean values of lift coefficient for a circular cylinder in orbital motion in a uniform flow*, in: Proc. 8th International Conference on Flow-Induced Vibration, Paris, Vol. **2**, (2004), pp. 93–98.

[8] Baranyi, L.: *Lift and drag evaluation in translating and rotating non-inertial systems*, Journal of Fluids and Structures **20**(1) (2005) 25–34.

[9] Baranyi, L. and Lewis, R.I.: *Comparison of a grid-based CFD method and vortex dynamics predictions of low Reynolds number cylinder flows*, The Aeronautical Journal **110**(1103) (2006) 63–71.

[10] Baranyi, L.: *Energy transfer between an orbiting cylinder and moving fluid*, in: Proc. PVP2006-ICPVT-11 ASME Pressure Vessels and Piping Division Conference, Vancouver, Canada (2006), on CD ROM, pp. 1–10.

[11] Baranyi, L.: *Orbiting cylinder at low Reynolds numbers*, in: Proc. IUTAM Symposium on Unsteady Separated Flows and Their Control, Corfu, Greece (2007), on CD ROM, pp. 1–5.

[12] Baranyi L.: *Mozgó henger körüli lamináris áramlás vizsgálata* (Habilitációs tézisfüzet, Miskolci Egyetem, Miskolc, 2007).

[13] BARANYI, L.: *Numerical simulation of flow around an orbiting cylinder at different ellipticity values*, Journal of Fluids and Structures **24** (2008) 833–906.

[14] BARKLEY, D. AND HENDERSON, R.D.: *Three-dimensional Floquet stability analysis of the wake of a circular cylinder*, Journal of Fluid Mechanics **322** (1996) 215–241.

[15] BEARMAN, P. W.: *Developments in the understanding of bluff body flows*, in: Proc. JSME Centennial Grand Congress, International Conference on Fluid Engineering, Vol. **1**, (1997), pp. 53–61.

[16] BLACKBURN, H.M. AND HENDERSON. R.D.: *A study of two-dimensional flow past an oscillating cylinder*, Journal of Fluid Mechanics **385** (1999) 255–286.

[17] BLACKBURN, H.M.: *"Computational bluff body fluid dynamics and aeroelasticity"*, in Coupling of Fluids, Structures and Waves Problems in Aeronautics, Notes on Numerical Fluid Mechanics and Multidisciplinary Design (NNFM), Vol. **85**, Eds. Barton, N.G. and Periaux, J. (Springer, Berlin, 2003) 10–23.

[18] BRONSTEJN, I.N., SZEMENGYAJEV, K.A., MUSIOL, G. ÉS MÜHLIG, H.: *Matematikai Kézikönyv* (8. kiadás, TypoTEX Kiadó, Budapest, 2002).

[19] DIDIER, E. AND BORGES, A.R.J.: *Numerical predictions of low Reynolds number flow over an oscillating circular cylinder*, Journal of Computational and Applied Mechanics, (in press).

[20] DYKE, M.V.: *An Album of Fluid Motion* (The Parabolic Press, Stanford, 1982).

[21] FARKAS M.: *Matematikai Kislexikon* (Műszaki Könyvkiadó, Budapest, 1979).

[22] FLETCHER C.A.J.: *Computational Techniques for Fluid Dynamics*, Vol. **2**, (Springer, Berlin, 1997).

[23] HARLOW, F.H. AND WELCH, J.E.: *Numerical calculation of time-dependent viscous incompressible flow of fluid with free surface*, Physics of Fluids **8** (1965) 2182–2189.

[24] KAWAMURA, T. AND KUWAHARA, K.: *Computation of high Reynolds number flow around a circular cylinder with surface roughness*, in: Proc. 22nd Aerospace Sci. Meeting, AIAA-84-0340, (1984), pp. 1–11.

[25] KOIDE, M., TOMIDA, S., TAKAHASHI, T., BARANYI, L. AND SHIRAKASHI, M.: *Influence of cross-sectional configuration on the synchronization of Kármán vortex shedding with the cylinder oscillation*, JSME International Journal, Series B **45**(2) (2002) 249–258.

[26] LU, X.Y. AND DALTON, C.: *Calculation of the timing of vortex formation from an oscillating cylinder*, Journal of Fluids and Structures **10** (1996) 527–541.

[27] MOON, F.C.: *Chaotic and Fractal Dynamics* (Wiley, New York, 1992).

[28] MUREITHI, N.W., COTOI, I. AND RODRIGUEZ, M.: *Response of the Karman wake to external periodic forcing and implications for vortex shedding control*, in: Proceedings of the 8th International Conference on Flow-Induced Vibration, Vol. **1**, (2004), pp. 87–92.

[29] NAYFEH, A.H. AND BALACHANDRAN, B.: *Applied Nonlinear Dynamics* (Wiley, New York, 1995).

[30] NICOLSKY, S.M.: *A Course of Matematical Analysis* , Vol. **2**, (Mir Publishers, Moscow, 1987).

[31] NORBERG, C.: *Fluctuating lift on a circular cylinder: review and new measurements*, Journal of Fluids and Structures **17** (2003) 57–96.

[32] PONCET, P.: *Vanishing of B mode in the wake behind a rotationally oscillating cylinder*, Physics of Fluids **14** (2002) 2021–2023.

[33] ROSHKO, A.: *Perspectives on bluff body aerodynamics*, Journal of Wind Engineering and Industrial Aerodynamics **49** (1993) 79–100.

[34] SCHLICHTING, H.: *Boundary-Layer Theory* (McGraw-Hill, New York, 1979).

[35] STANSBY, P.K. AND RAINEY, R.C.T.: *On the orbital response of a rotating cylinder in a current*, Journal of Fluid Mechanics **439** (2001) 87–108.

[36] THOMPSON, M.C. AND P. LE GAL: *The Stuart-Landau model applied to wake transition revisited*, European Journal of Mechanics B/Fluids **23** (2004) 219–228.

[37] UJVÁROSI, S. AND BARANYI, L.: *Numerical visualisation of cylinder flows*, in: Proc. Micro-CAD 2006, International Computer Science Conference, Miskolc, Section E, (2006), pp. 67–72.

[38] WILLIAMSON, C.H.K. AND ROSHKO, A.: *Vortex formation in the wake of an oscillating cylinder*, Journal of Fluids and Structures **2** (1988) 355–381.

[39] WILLIAMSON, C.H.K.: *Vortex dynamics in the cylinder wake*, Annual Review of Fluid Mechanics **28** (1996) 477–539.

BARANYI LÁSZLÓ
Miskolci Egyetem
Áramlás- és Hőtechnikai gépek Tanszéke
3515 Miskolc-Egyetemváros
E-mail: arambl@uni-miskolc.hu

NUMERICAL SIMULATION OF LOW REYNOLDS NUMBER FLOW AROUND AN ORBITING CYLINDER

LÁSZLÓ BARANYI

This paper deals with the 2D numerical simulation of low Reynolds number incompressible fluid flow around a circular cylinder in orbital motion placed in an otherwise uniform stream. Sudden changes (jumps) in state are found when time-mean or *rms* values of force coefficients – lift, drag, and base pressure coefficients – or energy transfer are plotted against ellipticity of the cylinder path. Pre- and post-jump analysis was carried out by investigating limit cycles, time histories, phase angles and flow patterns. These investigations revealed that ellipticity can have a large effect on the energy transfer between the fluid and a cylinder forced to follow an orbital path, and that small changes in the amplitude of transverse motion can have a major effect on the force coefficients. Phase angle between lift and transverse displacement changed by about 180° with the jumps. What triggers these changes is uncertain; probably there are two attractors, each with its 'basin of attraction' of this nonlinear system and the solution is attracted to one or the other of the attractors depending on the values of the parameters.

Alkalmazott Matematikai Lapok **26** (2009), 255-272.

OPERÁTORSZELETELÉSI ELJÁRÁSOK ÉS VIZSGÁLATUK

FARAGÓ ISTVÁN

Az operátorszeletelés (angolul "operator splitting") egy olyan széleskörűen elterjedt és sikeresen alkalmazott eljárás, amelynek segítségével a bonyolult szerkezetű feladatokat egyszerűbb feladatok sorozatára vezetjük vissza. Ebben a dolgozatban ismertetjük a legfontosabb operátorszeletelési eljárásokat, kitérve azok algoritmikus realizálásának kérdéseire is. Megvizsgáljuk az eljárás numerikus viselkedését, továbbá a szeletelt feladatok numerikus megoldása esetén az így nyert módszereket. Végezetül kapcsolatot teremtünk ezen kombinált módszerek és számos, az irodalomban ismeretes eljárás között.

1. Bevezetés

A matematikai modellalkotás során az összetett, időfüggő fizikai folyamatok folytonos modelljeként gyakran olyan parabolikus típusú parciális differenciálegyenleteket hozunk létre, amelyeknek stacionárius (időtől nem függő) elliptikus része egyszerűbb struktúrájú operátorok összegeként áll elő.

1.1. Példa. Tekintsük a d darab szennyezőanyag terjedését leíró légszennyeződési folyamat matematikai modelljét [18]:

$$\frac{\partial c_i}{\partial t} = -\nabla \cdot (\mathbf{u}c_i) + \nabla \cdot (\mathbf{K}\nabla c_i) - \sigma_i c_i + g_i(\mathbf{x}, t) + R_i(\mathbf{x}, c_1, \ldots, c_d). \tag{1}$$

A fenti képletben $i = 1, 2, \ldots d$, és $c_i = c_i(\mathbf{x}, t)$ jelöli az i-edik szennyezőanyag koncentrációját. A képlet jobb oldalán szereplő tagok rendre az egyes fizikai részfolyamatokat leíró tagok, nevezetesen, az advekció, a turbulens diffúzió, az ülepedés, a szennyezőanyag kibocsátása és a kémiai reakciók. Feltételezzük, hogy a koncentráció-eloszlás a kezdeti időpontban ($t = 0$) ismert.

A továbbiakban az ilyen típusú feladatokat egységesen, ún. operátoralakban a következő módon adjuk meg:

$$\begin{cases} \dfrac{dw(t)}{dt} = Aw(t) \equiv \displaystyle\sum_{i=1}^{d} A_i w(t), \quad t \in (0, T) \\ w(0) = w_0, \end{cases} \tag{2}$$

ahol w az ismeretlen függvény, w_0 adott elem, A_i, $(i = 1, 2, \ldots d)$ adott operátorok. (Megjegyezzük, hogy amennyiben vannak peremfeltételek, akkor azok az operátorok értelmezési tartományában szerepelnek.)

A (2) feladat (az ún. absztrakt Cauchy-feladat) pontos megoldása speciális esetben formálisan közvetlenül is felírható. Amikor A olyan lineáris operátor, amely egy C_0-félcsoportot generál, akkor

$$w(t) = \exp(tA)w(0), \tag{3}$$

ahol $\exp(tA)$ az A operátor által generált félcsoport. (Ha az A operátor korlátos, akkor $\exp(tA)$ előállítása az exp függvény sorából közvetlen behelyettesítéssel adódik. A további részleteket lásd [7] könyvben.)

Mivel a fenti megoldás előállítása (ha az egyáltalán lehetséges) többnyire csak formális, ezért a numerikus módszerek alkalmazása többnyire elkerülhetetlen. Ennek lényege, hogy az exponenciális függvényt approximáljuk valamilyen (általában) racionális függvénnyel, azaz

$$\exp(z) \sim r(z). \tag{4}$$

Ekkor a numerikus módszer algoritmusa

$$y^{n+1} = r(\tau A)y^n, \tag{5}$$

ahol $\tau > 0$ a diszkretizációs paraméter, és y^n jelenti az approximációt a $t = n\tau$ időrétegen.

1.2. Példa. Legyen

$$r(z) = \frac{1 + (1-\theta)z}{1 - \theta z}$$

az ún. θ-módszer, és $d = 2$. Ekkor:

$$y^{n+1} = r_\theta(\tau(A_1 + A_2))y^n,$$

ahol

$$r_\theta(\tau(A_1 + A_2)) = (\, I - \theta\tau(A_1 + A_2))^{-1}(\, I + (1-\theta)\tau(A_1 + A_2))$$

és I az identitásoperátor.

A fenti megközelítés hiányossága, hogy nem használjuk fel az A operátor speciális szerkezetét, nevezetesen, hogy több, egyszerűbb struktúrájú operátor összege. A továbbiakban célunk olyan eljárás definiálása, amely alkalmas a fenti sajátosság kihasználására.

Megjegyzés. Az (5) módszer ténylegesen időbeli diszkretizálást jelent az $\omega_\tau = \{t_n = n\tau, n = 0, 1, \ldots N; N\tau = T\}$ rácshálón. Így ha parciális differenciálegyenletekre közvetlenül alkalmazzuk, akkor (5) időrétegenként egy-egy elliptikus

típusú feladat megoldását jelenti, azaz további numerikus módszer alkalmazása szükséges. Ugyanakkor, ha (5) alkalmazása előtt már térben diszkretizáljuk a feladatunkat (pl. véges differenciák vagy véges elemek módszerével), és A_h jelenti a diszkretizált operátort, akkor az így nyert

$$y^{n+1} = r(\tau A_h)y^n \tag{6}$$

egylépéses iteráció már közvetlenül alkalmas a számítások elvégzésére. Ekkor ugyanis A_h mátrixként reprezentálható, és így (6) időrétegenként lineáris algebrai egyenletrendszerek megoldását jelenti.

2. Operátorszeletelés

Ebben a szakaszban feltételezzük, hogy $A_i : \mathbf{X} \to \mathbf{X}$ egész téren értelmezett lineáris operátorok, ahol $\mathbf{X}$ valamely rögzített Banach-tér, és operátornormán az indukált szuprémum normát értjük. A (2) absztrakt Cauchy-feladat megoldásának azon $w : (0,T) \to \mathbf{X}$ függvényt nevezzük, amely folytonosan differenciálható $(0,T)$-n és kielégíti a (2) feladatot. Így az $\mathbf{X}$ tér megválasztásától függően egyaránt tárgyalható a klasszikus, illetve a gyenge megoldás. (Ha gyenge megoldásról beszélünk, akkor az előforduló deriváltak is általánosított értelemben értendők és ez a tárgyalásmódot bonyolultabbá teheti.) Fontos megjegyeznünk, hogy az időfüggő parciális differenciálegyenletek esetén általában nem teljesül az A_i operátorokra tett feltételezésünk, mint például (1) feladatban sem. Ilyenkor szokásos módon az A_i operátorok a már térben diszkretizált (továbbá, nemlineáris operátorok esetén a már linearizált) operátorokat jelentik, azaz (2) absztrakt Cauchy-feladat ténylegesen a szemidiszkretizált lineáris feladatot jelenti.

Megjegyezzük, hogy általában homogén peremfeltételeket tételezünk fel. Amennyiben a peremfeltétel inhomogén, akkor a szokásos módon homogenizálhatjuk a feladatot, illetve a szemidiszkrét feladatot ennek figyelembe vételével írhatjuk fel.

Amikor a (4) típusú approximációt alkalmazzuk a (2) feladatra, akkor valójában az $\exp\left(\sum_1^d z_i\right)$ függvényt kell közelítenünk. Ennek egy lehetséges módja a (4) approximáció, amikor is előbb approximáljuk az exponenciális függvényt, és utána helyettesítjük be összegként az operátort. Ezzel természetesen nem tudjuk kihasználni az operátor speciális alakját.

Az operátorszeletelés alapötlete, hogy a (4) approximáció során az $\exp\left(\sum_1^d z_i\right)$ függvényt első lépésben nem racionális, hanem az egyes részoperátorok exponenciálisainak segítségével approximáljuk.

2.1. Példa. A legkézenfekvőbb az

$$\exp\left(\sum_{i=1}^{d} z_i\right) \sim \prod_{i=1}^{d} \exp(z_i) \tag{7}$$

típusú közelítés.

2.2. Példa. Bevezethetjük az

$$\exp\left(\sum_{i=1}^{d} z_i\right) \sim \prod_{i=1}^{d-1} \exp\left(\frac{z_i}{2}\right) \exp(z_d) \prod_{i=1}^{d-1} \exp\left(\frac{z_{d-i}}{2}\right) \tag{8}$$

típusú közelítést is.

2.3. Példa. Vegyük észre, hogy a (7) közelítésben lényeges a sorrend a jobb oldalon. Ezért célszerűnek látszik az

$$\exp\left(\sum_{i=1}^{d} z_i\right) \sim \frac{1}{2}\left[\prod_{i=1}^{d} \exp(z_i) + \prod_{i=1}^{d} \exp\left(z_{d+1-i}\right)\right] \tag{9}$$

típusú közelítés is.

Nyilvánvalóan, skalárok esetén (7), (8) és (9) egyenlőséget jelent, de tetszőleges korlátos operátorokra az egyenlőség nem áll fenn. Ugyanakkor, ha az operátorok páronként kommutálnak, akkor ismételten igaz az egyenlőség.

A fenti közelítéseket alkalmazhatjuk a (2) feladat közelítő megoldására a következő módon. Tegyük fel, hogy A_i operátorok szintén generátorok, és vezessük be a következő operátorfüggvényeket:

$$r_{\text{szek}}(A) := \prod_{i=1}^{d} \exp(A_i);$$

$$r_{\text{SM}}(A) := \prod_{i=1}^{d-1} \exp\left(\frac{1}{2}A_i\right) \exp(A_d) \prod_{i=1}^{d-1} \exp\left(\frac{1}{2}A_{d-i}\right)$$

és

$$r_{\text{szim}}(A) := \frac{1}{2}\left[\prod_{i=1}^{d} \exp(A_i) + \prod_{i=1}^{d} \exp(A_{d+1-i})\right].$$

Ezen operátorok segítségével definiálhatók az új numerikus módszerek, nevezetesen

$$w^N_{\text{szel}}((n+1)\tau) = r_{\text{szel}}(\tau A) w^N_{\text{szel}}(n\tau), \quad n = 0, 1, \ldots N, \tag{10}$$

ahol szel $\in \{\text{szek}; SM; \text{szim}\}$ és $w^N_{\text{szel}}(n\tau)$ az adott szeleteléshez tartozó numerikus megoldás az ω_τ rácshálón. A fenti operátorszeletelési eljárásokat rendre *szekvenciális, Strang–Marcsuk* és *szimmetrikus szekvenciális szeleteléseknek* nevezzük.

A (10) algoritmus realizálásának lényeges pontja az $\exp(A_i)$ kiszámítása, pontosabban, $\exp(\tau A_i)v$ meghatározása valamely adott v elem esetén. Mivel ez nem más, mint egy τ hosszúságú intervallumon értelmezett, v kezdeti vektorú, A_i operátorú homogén Cauchy-feladat megoldása, ezért a fenti módszerek algoritmikus relizálása a következő.

1. *Szekvenciális szeletelés [1].* Valamennyi rögzített $n = 1, 2, \dots N$ értékre rendre megoldjuk a következő d darab Cauchy-feladatot:

$$\frac{dw_i^n}{dt}(t) = A_i w_i^n(t), \qquad (n-1)\tau < t \le n\tau,$$
$$w_i^n((n-1)\tau) = w_{i-1}^n(n\tau),$$

ahol $i = 1, 2, \dots d$. A szeletelt megoldás

$$w_{\text{szek}}^N(n\tau) = w_d^n(n\tau)$$

és az algoritmusban $w_0^n(n\tau) = w_{\text{szek}}^N((n-1)\tau)$, továbbá $w_{\text{szek}}^N(0) = w(0)$ a (2) kezdeti feltételből ismert w_0 elem. Tehát az algoritmus:

$$\underbrace{A_1 \to A_2 \to \cdots A_d}_{\text{1. lépés}} \Rightarrow \underbrace{A_1 \to A_2 \to \cdots A_d}_{\text{2. lépés}} \Rightarrow \cdots \Rightarrow \underbrace{A_1 \to A_2 \to \cdots A_d}_{\text{N. lépés}}.$$

2. *Strang–Marcsuk-szeletelés [16], [13]* . A rögzített $n = 1, 2, \dots N$ értékekre rendre megoldjuk a következő, összességében $2d-1$ darab Cauchy-feladatot.

Először $i = 1, 2, \dots d-1$ értékekre megoldjuk a

$$\frac{dw_i^n}{dt}(t) = A_i w_i^n(t), \qquad (n-1)\tau < t \le (n-0,5)\tau,$$
$$w_i^n((n-1)\tau) = w_{i-1}^n((n-0,5)\tau)$$

feladatokat. Ezután megoldjuk a

$$\frac{dw_d^n}{dt}(t) = A_d w_d^n(t), \qquad (n-1)\tau < t \le n\tau,$$
$$w_d^n((n-1)\tau) = w_{d-1}^n((n-0,5)\tau)$$

feladatot. A rögzített n mellett végezetül $i = d+1, d+2, \dots 2d-1$ értékekre megoldjuk a

$$\frac{dw_i^n}{dt}(t) = A_{2d-i} w_i^n(t), \qquad (n-0.5)\tau < t \le n\tau,$$
$$w_i^n((n-0.5)\tau) = w_{i-1}^n(n\tau)$$

feladatokat.

A szeletelt megoldás

$$w_{\mathrm{SM}}^N(n\tau) = w_{2d-1}^n(n\tau)$$

és az algoritmusban

$$w_0^n((n-0,5)\tau) = w_{\mathrm{SM}}^N((n-1)\tau),$$

továbbá

$$w_{\mathrm{SM}}^N(0) = w(0)$$

a (2) kezdeti feltételből ismert w_0 elem. Tehát az algoritmus:

$$\underbrace{\frac{1}{2}A_1 \to \frac{1}{2}A_2 \to \cdots \frac{1}{2}A_{d-1}}_{\text{1a. lépés}} \to \underbrace{A_d}_{\text{1b. lépés}} \to \underbrace{\frac{1}{2}A_{d-1} \to \frac{1}{2}A_{d-2} \to \cdots \frac{1}{2}A_1}_{\text{1c. lépés}} \Rightarrow$$

$$\cdots\cdots\cdots$$

$$\Rightarrow \underbrace{\frac{1}{2}A_1 \to \frac{1}{2}A_2 \to \cdots \frac{1}{2}A_{d-1}}_{\text{Na. lépés}} \to \underbrace{A_d}_{\text{Nb. lépés}} \to \underbrace{\frac{1}{2}A_{d-1} \to \frac{1}{2}A_{d-2} \to \cdots \frac{1}{2}A_1}_{\text{Nc. lépés}}.$$

3. *Szimmetrikus szekvenciális szeletelés [15], [5].* Rögzített $n = 1, 2, \ldots N$ értékekre rendre megoldjuk a következő $2d$ darab Cauchy-feladatot:

$$\frac{dv_i^n}{dt}(t) = A_i v_i^n(t), \qquad (n-1)\tau < t \le n\tau,$$
$$v_i^n((n-1)\tau) = v_{i-1}^n(n\tau),$$

és

$$\frac{du_i^n}{dt}(t) = A_{d+1-i} u_i^n(t), \qquad (n-1)\tau < t \le n\tau,$$
$$u_i^n((n-1)\tau) = u_{i-1}^n(n\tau),$$

ahol $i = 1, 2, \ldots d$. A szeletelt megoldás

$$w_{\mathrm{szim}}^N(n\tau) = \frac{v_d^n(n\tau) + u_d^n(n\tau)}{2}.$$

A fenti algoritmusban

$$v_0^n(n\tau) = u_0^n(n\tau) = w_{\mathrm{szek}}^N((n-1)\tau),$$

továbbá

$$w_{\mathrm{szim}}^N(0) = w(0)$$

a (2) kezdeti feltételből ismert w_0 elem.

Tehát az algoritmus:

$$\left.\begin{array}{r} A_1 \to A_2 \to \cdots A_d \\ A_d \to A_{d-1} \to \cdots A_1 \\ \text{(1. lépés)} \end{array}\right\} \Longrightarrow \cdots \Longrightarrow \left.\begin{array}{r} A_1 \to A_2 \to \cdots A_d \\ A_d \to A_{d-1} \to \cdots A_1 \\ \text{(N. lépés)} \end{array}\right\}.$$

Megjegyezzük, hogy a fenti három szeletelési eljárás mellett még számos egyéb, sikeresen alkalmazott eljárás is létezik. Ugyanakkor dokgozatunkban a továbbiakban mi a felsoroltakat (elsősorban azok elterjedtsége miatt) elemezzük részletesebben.

3. A lokális szeletelési hiba

Az operátorszeletelés, mint a numerikus módszerek általában, nem eredményeznek pontos megoldást, azaz tetszőleges operátorok esetén a szeletelt megoldás nem egyezik meg a (2) feladat pontos megoldásával az ω_τ rácsháló pontjaiban. Az első időlépés utáni eltérést kiemelten kezelve vezessük be a következő meghatározást.

3.1. Definíció. Az $Err_{\text{szel}} = w(\tau) - w^N_{\text{szel}}(\tau)$ kifejezést az adott szeletelési eljárás *lokális szeletelési hibájának* nevezzük.

Természetes elvárás, hogy τ nullához tartása mellett a lokális szeletelési hiba is nullához tartson. Mivel ezen konvergencia rendje jellemzi a szeletelés minőségét, célszerű bevezetni az alábbi definíciót.

3.2. Definíció. Azt mondjuk, hogy az adott szeletelési eljárás p-ed rendű, ha $Err_{\text{szel}} = \mathcal{O}\left(\tau^{p+1}\right)$. Egy adott szeletelést konzisztensnek nevezünk, ha rendje $p > 0$.

A továbbiakban korlátos operátorok esetén meghatározzuk az egyes szeletelések rendjét.

3.1. Példa. Vizsgáljuk meg a szekvenciális szeletelés rendjét! A (3) összefüggés alapján a pontos megoldás a $t = \tau$ pontban

$$w(\tau) = \exp(\tau A)w(0) = \sum_{n=0}^{\infty} \frac{1}{n!}\tau^n A^n w_0 = \left(I + \tau A + \frac{1}{2}\tau^2 A^2\right) w_0 + \mathcal{O}(\tau^3).$$

Figyelembe véve, hogy $A = \sum_{i=1}^{d} A_i$, könnyen láthatón

$$w(\tau) = \left(I + \tau \sum_{i=1}^{d} A_i + \frac{\tau^2}{2} \sum_{i,j=1}^{d} A_i A_j\right) w_0 + \mathcal{O}(\tau^3). \tag{11}$$

A szekvenciálisan szeletelt közelítő megoldás az első lépés után

$$w^N_{\text{szek}}(\tau) = \prod_{i=1}^{d} \exp(\tau A_i) w_0 = \prod_{i=1}^{d} \left(I + \tau A_i + \frac{\tau^2}{2} A_i^2 \right) w_0 + \mathcal{O}(\tau^3).$$

Ezért tehát

$$w^N_{\text{szek}}(\tau) = \left(I + \tau \sum_{i=1}^{d} A_i + \frac{\tau^2}{2} \sum_{i=1}^{d} A_i^2 + \tau^2 \prod_{\substack{i,j=1 \\ i<j}}^{d} A_i A_j \right) w_0 + \mathcal{O}(\tau^3). \tag{12}$$

A (11) és (12) képletek alapján a lokális szeletelési hiba tehát

$$\begin{aligned} Err_{szek} &= \frac{\tau^2}{2} \left(\prod_{\substack{i,j=1 \\ i>j}}^{d} A_i A_j - \prod_{\substack{i,j=1 \\ i<j}}^{d} A_i A_j \right) w_0 + \mathcal{O}(\tau^3) = \\ &= \frac{\tau^2}{2} \prod_{\substack{i,j=1 \\ i>j}}^{d} (A_i A_j - A_j A_i) w_0 + \mathcal{O}(\tau^3). \end{aligned} \tag{13}$$

Mivel általános esetben a jobb oldal $\mathcal{O}(\tau^2)$, ezért a szekvenciális szeletelés elsőrendű módszer.

3.2. Példa. Vizsgáljuk meg a szimmetrikus szekvenciális szeletelés rendjét! Az előző számítások alapján nyilvánvaló, hogy ebben az esetben

$$\begin{aligned} w^N_{\text{szim}}(\tau) &= \left(I + \tau \sum_{i=1}^{d} A_i + \frac{\tau^2}{2} \sum_{i=1}^{d} A_i^2 + \frac{\tau^2}{2} \prod_{\substack{i,j=1 \\ i<j}}^{d} (A_i A_j + A_j A_i) \right) w_0 + \\ &+ \mathcal{O}(\tau^3) = \left(I + \tau \sum_{i=1}^{d} A_i + \frac{\tau^2}{2} \sum_{i,j=1}^{d} A_i A_j \right) w_0 + \mathcal{O}(\tau^3). \end{aligned} \tag{14}$$

A (11) és (14) képletek alapján a lokális szeletelési hiba tehát

$$Err_{\text{szek}} = \mathcal{O}(\tau^3),$$

azaz a szimmetrikus szekvenciális szeletelés másodrendű.

Megjegyezzük, hasonló módon megmutatható, hogy a Strang–Marcsuk-szeletelés is másodrendű. (Ennek belátását az Olvasóra bízzuk.)

4. A lokális szeletelési hiba elemzése

A továbbiakban a szeletelések lokális hibájára vonatkozóan teszünk néhány megjegyzést.

1. Fontos kérdés a lokális szeletelési hiba eltűnésének vizsgálata, azaz megadni azt a feltételt, amely mellett a szeletelt megoldás és a pontos megoldás megegyezik. A lokális szeletelési hiba megjelenésének oka, hogy általában $\exp(A_1 + A_2) \neq \exp(A_1)\exp(A_2)$. Mint ismeretes, az egyenlőség csak akkor érvényes, amikor az operátorok kommutálnak, azaz az $[A_1, A_2] := A_1A_2 - A_2A_1$ kommutátorra $[A_1, A_2] = 0$ érvényes. Ez egyrészt összhangban van a szekvenciális szeletelés lokális hibájára nyert (13) kifejezéssel: a másodrendű hibatag eltűnik, ha mindegyik operátorpár kommutál. Emellett azt is jelenti, hogy nem csak a másodrendű, hanem valamennyi rendű hibatag eltűnik.

2. Az előző észrevétel azt jelenti, hogy páronként kommutáló operátorok esetén a szeletelések pontosak. (Természetesen feltételezve, hogy a szeletelt rész-feladatokat pontosan, azaz numerikus módszer alkalmazása nélkül oldjuk meg.) Felmerül a kérdés: vajon a páronkénti kommutálás szükséges feltétele-e a lokális szeletelési hiba eltűnésének?

   A következő egyszerű példa választ ad erre a kérdésre [10].

   4.1. Példa. Tekintsük a következő 2×2-es mátrixot:

$$A = \begin{bmatrix} 4 & 2 \\ 0 & 3 \end{bmatrix}.$$

   Bontsuk fel három mátrix összegére, azaz írjuk fel $A_1 + A_2 + A_3$ alakban, ahol

$$A_1 = A_3 = \begin{bmatrix} 3 & 1 \\ 0 & 2 \end{bmatrix} \quad \text{és} \quad A_2 = \begin{bmatrix} -2 & 0 \\ 0 & -1 \end{bmatrix}.$$

   Ekkor

$$e^{tA} = \begin{bmatrix} e^{4t} & 2e^{3t}(e^t - 1) \\ 0 & e^{3t} \end{bmatrix},$$

$$e^{tA_1} = e^{tA_3} = \begin{bmatrix} e^{3t} & e^{2t}(e^t - 1) \\ 0 & e^{2t} \end{bmatrix}, \quad \text{és} \quad e^{tA_2} = \begin{bmatrix} e^{-2t} & 0 \\ 0 & e^{-t} \end{bmatrix}.$$

   Ebben a példában A_1 és A_2 nem kommutálnak, mivel

$$[A_1, A_2] = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

   Ugyanakkor,

$$e^{\tau A_1} e^{\tau A_2} e^{\tau A_3} = e^{\tau A},$$

ami azt jelenti, hogy a szekvenciális szeletelés pontos. Tehát a szekvenciális szeletelés esetén a páronkénti kommutálás nem szükséges feltétele a lokális szeletelési hiba eltűnésének. Megjegyezzük, hogy $d = 2$ (azaz két operátor) esetén a lokális szeletelési hiba (v.ö. (13) képlettel):

$$Err_{\text{szek}} = \frac{\tau^2}{2}\left(A_1A_2 - A_2A_1\right)w_0 + \mathcal{O}(\tau^3) = \frac{\tau^2}{2}[A_1, A_2]w_0 + \mathcal{O}(\tau^3), \quad (15)$$

amely azt mutatja, hogy két operátor esetén a kommutálás szükséges feltétel is.

Mivel $d = 2$ esetén könnyen láthatóan a Strang–Marcsuk-szeletelés nem más, mint egy három operátoros szekvenciális szeletelés (ahol az operátorok rendre $A_1/2$, A_2 és $A_1/2$), az előzőekből következik, hogy a páronkénti kommutálás a Strang–Marcsuk-szeletelés esetén sem szükséges feltétel. (Hasonlóan belátható az állítás a szimmetrikus szekvenciális szeletelésre is.)

3. Egy adott numerikus módszer esetén ismeretes fogalom a konzisztencia. Érdekes kapcsolat áll fenn az operátorszeletelés konzisztenciája (lásd 3.2. Definíció), illetve az operátorszeletelésnek mint numerikus módszernek a konzisztenciája között. A numerikus módszer konzisztenciájához az szükséges, hogy bármilyen időpontban is vesszük a pontos megoldást, onnan egy időlépést téve a numerikus módszerrel, a pontos megoldástól való eltérés (lokális hiba) nullához tarson. Vagyis azt követeljük meg, hogy

$$\sup_t \frac{1}{\tau}\left(r_{\text{szel}}(\tau A)u(t) - \exp(\tau A)u(t)\right) \quad (16)$$

tartson nullához $\tau \to 0$ esetén. Mivel $t = 0$-ban ez éppen a lokális szeletelési hiba nullához tartását jelenti, ezért a szeletelés mint numerikus módszer konzisztenciájából következik a lokális szeletelési hiba nullához tartása, és így a szeletelés 3.2. Definíció szerinti konzisztenciája is. Ugyanakkor ez megfordítva nem feltétlenül érvényes.

4. Fontos kitérnünk a nemkorlátos operátorok konzisztenciájának vizsgálatára, természetesen továbbra is feltéve, hogy ezek az operátorok C_0-félcsoportok generátorai. Bár ekkor nem értelmezhető a félcsoport a generátorának hatványsoros előállításával, a Taylor-sor maradéktaggal való megadásával az analízis mégis elvégezhető. Megmutatható, hogy kellően sima kezdeti függvény esetén a korlátos operátorokra kimutatott rend ebben az esetben is érvényben marad. (A szekvenciális szeletelésre a [2], míg a Strang–Marcsuk, illetve a szimmetrikus szekvenciális szeletelésekre a [9] cikkben található a bizonyítás.)

5. A lokális szeletelés hibájában a rend azt fejezi ki, hogy megfelelően kicsiny τ mellett érvényes csak a becslés. Például $d = 2$ esetén a szekvenciális szeletelés hibájából (lásd (15)) arra következtethetnénk, hogy általában

a kommutátor normája egyértelműen meghatározza a lokális hiba nagyságát, a τ szeletelési lépésköz megválasztásától függetlenül. A következő példa [8] megmutatja, hogy ez általában nem érvényes.

4.2. Példa. Legyenek

$$A_1 = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{3}{4} \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$A = A_1 + A_2$, $p \in \mathbb{R}$ tetszőleges, nemnulla állandó, $B_1 = pA_1$, $B_2 = \frac{1}{p}A_2$, $B = B_1 + B_2$, $e = (1,1,1,1)$ és $\tau = 1$. Tekintsük a következő problémákat:

$$\left.\begin{aligned} u'(t) &= Au(t), \quad t \in (0,1] \\ u(0) &= e \end{aligned}\right\} \tag{17}$$

és

$$\left.\begin{aligned} w'(t) &= Bw(t), \quad t \in (0,1] \\ w(0) &= e \end{aligned}\right\}. \tag{18}$$

Könnyen látható, hogy

$$\|[A_1, A_2]\|_\infty = \|[B_1, B_2]\|_\infty = \frac{1}{6},$$

azaz a kommutátorok normája megegyezik.

Legyen most $p = 1000$, és alkalmazzuk a szekvenciális szeletelést a (17) és a (18) feladatokra. Egyszerű számolással ellenőrizhető, hogy a lokális szeletelési hibák

$$\left\|Err^A_{Sp}(\tau = 1)\right\|_\infty = \left\|e^{(A_1+A_2)} - e^{A_2}e^{A_1}\right\|_\infty = 0,125$$

és

$$\left\|Err^B_{Sp}(\tau = 1)\right\|_\infty = \left\|e^{(B_1+B_2)} - e^{B_2}e^{B_1}\right\|_\infty = 20,8334.$$

Látható, hogy a második esetben lényegesen nagyobb (166,67-szoros) hibát kaptunk. Megjegyezzük, hogy p növelésével ez az eltérés tetszőlegesen naggyá tehető. A jelenség oka, hogy a fenti τ mellett a magasabb rendű tagok szerepe még jelentős. Természetesen τ megfelelő csökkentésével a kommutátor normája már meghatározza a lokális szeletelési hibát.

6. Felmerülhet a kérdés: ha a (2) homogén egyenlet helyett a

$$\begin{cases} \dfrac{dw(t)}{dt} = Aw(t) + f(t) \equiv \sum_{i=1}^{d} A_i w(t) + \sum_{i=1}^{d} f_i(t), & t \in (0,T) \\ w(0) = w_0, \end{cases}$$

inhomogén feladatot tekintjük, akkor lehetséges-e a rend megtartásával szeletelni a feladatokat? Például a

$$\begin{aligned} &\frac{dw_i^n}{dt}(t) = A_i w_i^n(t) + f_i(t), \qquad (n-1)\tau < t \le n\tau, \\ &w_i^n((n-1)\tau) = w_{i-1}^n(n\tau), \end{aligned}$$

szekvenciális szeletelés (avagy az értelemszerűen átfogalmazott Strang–Marcsuk, illetve szimmetrikus szekvenciális szeletelések) esetén megőrződik-e a homogén feladatra korábban belátott rend? A válasz pozitív, azaz megfelelően sima forrástagok ($f_i(t)$) esetén megmarad a rend. (Lásd [4].)

7. Mi az eddigiekben a lokális hibát vizsgáltuk. Ugyanakkor a gyakorlat szempontjából a globális hiba fontos, vagyis az, hogy egy rögzített $t \in (0,T)$ pontban a $w_{\text{szel}}^N(n\tau)$ szeletelt megoldás (ahol $n\tau = t$, azaz $\tau \to 0$ miatt $n \to \infty$) tart-e, és ha igen, milyen rendben a $w(t)$ pontos megoldáshoz? A válasz megadása előtt az operátorszeletelésre mint numerikus eljárásra megfogalmazzuk a jól ismert stabilitásfogalmat.

4.1. Definíció. Egy operátorszeletelést stabilnak nevezünk, ha a (10) iterációban szereplő $r_{\text{szel}}(\tau A)$ operátorra az

$$\{\|r_{\text{szel}}^n(\tau A)\|, \quad n\tau \le t, \quad \tau > 0\}$$

operátorsereg egyenletesen korlátos.

A Lax-féle ekvivalenciatétel (pld. [12]) korrekt kitűzésű absztrakt Cauchy-feladatok esetén egy numerikus módszer esetén kapcsolatot teremt a konzisztencia, a stabilitás és a konvergencia között. Nevezetesen, konzisztencia esetén a stabilitás és a konvergencia ekvivalens egymással. Következésképpen, ha az operátorszeletelés stabil, akkor a globális hiba nullához tart. Emellett a konvergencia rendje általában p, azaz megegyezik a módszer rendjével. A gyakorlatban a stabilitás belátása nem könnyű, de kontraktív esetben, azaz amikor

$$\|r_{\text{szel}}(\tau A)\| \le 1,$$

a stabilitás nyilvánvalóan igaz. Így tehát ha az A_i operátorok kontrakciós félcsoportok generátorai, akkor a szeletelés mint numerikus módszer konvergens.

5. Numerikus módszerek alkalmazása a szeletelt feladatokra

Az eddigiekben azzal a feltételezéssel éltünk, hogy a szeleteléssel nyert egyes részfeladatokat pontosan oldjuk meg. Ugyanakkor ez általában nem lehetséges: bár feltételezéseink szerint az egyes A_i részoperátorok egyszerűbb struktúrájúak, a valós feladatokra nem reális feltételezni, hogy a szeletelt problémák pontosan megoldhatók. (Az egyszerűbb struktúra „csak" azt eredményezi, hogy a szeletelt feladatokra többnyire az irodalomból ismert, megbízható és hatékony numerikus eljárást tudunk alkalmazni.) Jelölje a továbbiakban Δt a szeletelt feladatra alkalmazott numerikus diszkretizáció lépésközét. (Nyilvánvalóan $\Delta t \leq \tau$.) Ez meghatároz egy $\omega_{\Delta t} = \{t_n = n\Delta t, n = 0, 1, \ldots N; N\Delta t = T\}$ rácshálót. Természetes elvárás, hogy $\omega_{\Delta t}$ a szeletelésre használt rácsháló finomítása legyen, azaz teljesüljön a $\omega_{\Delta t} \supset \omega_\tau$ tartalmazás. Ezért célszerű a $\Delta t = \tau/K$ megválasztást alkalmazni, ahol K adott természetes szám. ($K = 1$ esetén a szeletelési lépésköz és a numerikus lépésköz megegyezik, tehát mindegyik részfeladat numerikus megoldása során egyetlen lépést hajtunk végre.)

Az egyes szeletelt részfeladatokra valamely numerikus módszert alkalmazva a teljes feladatra egy globális diszkretizációt nyerünk az $\omega_{\Delta t}$ rácshálón. (Fontos kiemelni, hogy a választott numerikus módszerek eltérőek is lehetnek, hiszen gyakran éppen ez a szeletelés alkalmazásának célja.) Ezért a teljes diszkretizációs operátor függ a választott szeleteléstől, a szeletelési lépésköztől, az alkalmazott numerikus módszertől és ennek lépésközétől. Legyen például $d = 2$, és alkalmazzuk a τ lépésközű szekvenciális szeletelést. Válasszuk az első szeletelt feladatra az NM1 numerikus módszert a $\Delta t_1 = \tau/K_1$ lépésközzel, a második feladatra az NM2 numerikus módszert a $\Delta t_2 = \tau/K_2$ lépésközzel. Ekkor a teljes diszkretizációs operátor felírható $C_{\text{tot}} = C_{\text{tot}}(\tau, NM1, K_1, NM2, K_2)$ alakban. (Amikor $NM1, K_1, NM2$ és K_2 rögzítettek, akkor az egyszerűség kedvéért a továbbiakban a $C(\tau)$ jelölést alkalmazzuk.)

5.1. Példa. Legyen $d = 2$ a (2) feladatban, és alkalmazzuk a szekvenciális szeletelést. Válasszuk az explicit Euler (EE) módszert mindkét szeletelt feladatra $\Delta t = \tau$ megválasztással. (Azaz, $C_{\text{tot}} = C_{\text{tot}}(\tau, EE, 1, EE, 1)$). Ha y_1^n és y_2^n jelöli az $w_1^n(n\tau)$ és $w_2^n(n\tau)$ szeletelt pontos megoldások közelítéseit, akkor a numerikus séma a következő:

$$\frac{y_1^{n+1} - y_1^n}{\tau} = A_1 y_1^n, \quad \frac{y_2^{n+1} - y_2^n}{\tau} = A_2 y_2^n$$

és $y_2^n = y_1^{n+1}$. Ezért

$$y_2^{n+1} = (I + \tau A_2)(I + \tau A_1) y_1^n, \tag{19}$$

azaz a teljes diszkretizáló operátor

$$C(\tau) = (I + \tau A_2)(I + \tau A_1).$$

Korlátos operátorokra, korlátos időintervallumon könnyen megmutatható a fenti módszer konvergenciája.

5.1. ÁLLÍTÁS. *Tegyük fel, hogy a $[0,T]$ intervallumon értelmezett (2) feladatban ($d = 2$) az A_1 és A_2 korlátos operátorok. Ekkor a (19) módszer konvergens.*

Bizonyítás. Először a konzisztenciát látjuk be, azaz megmutatjuk, hogy $r_{\text{szel}}(\tau A) = C(\tau)$ megválasztással (16) érvényes. Mivel

$$\left\| \frac{1}{\tau} \left(C(\tau)u(t) - \exp(\tau A)u(t) \right) \right\| \leq \frac{1}{\tau} \|(C(\tau) - \exp(\tau A))\| \|u(t)\| =$$

$$= \left[\frac{\tau}{2} \|(A_1 - A_2)^2\| + \mathcal{O}(\tau^2) \right] \| \exp(tA) w_0 \| \leq$$

$$\leq \left[\frac{\tau}{2} \|(A_1 - A_2)^2\| + \mathcal{O}(\tau^2) \right] \exp(T\|A\|) \|w_0\| = const \cdot \tau + \mathcal{O}(\tau^2),$$

és az utolsó tag t-től függetlenül tart nullához $\tau \to 0$ esetén.

A stabilitás a

$$\|C(\tau)^n\| = \| \left((I + \tau A_2)(I + \tau A_1) \right)^n \| \leq (1 + \tau \|A_2\|)^n (1 + \tau \|A_1\|)^n \leq$$

$$\leq \exp(T\|A_2\|) \exp(T\|A_1\|) = const$$

relációból nyilvánvalóan következik. □

Megjegyezzük, hogy ha A_i nem korlátos (és az időintervallum sem az), de kontrakciós félcsoportot generál, akkor is érvényes marad az állítás. (A bizonyítás önállóan elvégezhető.) •

6. Néhány ismert eljárás mint kombinált módszer

Az 5. szakaszban definiáltuk a szeleteléssel és a numerikus módszer megválasztásával nyerhető teljes diszkretizációt.

Ebben a részben megmutatjuk, hogy a szeletelés a szeletelt feladatok numerikus megoldási módszerének alkalmas megválasztásával több, az irodalomból jól ismert módszer nyerhető.

6.1. A Crank–Nicolson-módszer

Tekintsük a

$$\begin{aligned} &\frac{dw}{dt} = Aw(t), \quad 0 < t \leq T \\ &w(0) = w_0 \end{aligned} \tag{20}$$

Cauchy-feladatot és alkalmazzuk a triviális $A = \frac{1}{2}A + \frac{1}{2}A$ szeletelést! Ekkor az (20)

feladatot a szekvenciális szeleteléssel diszkretizálva a következő szeletelt feladatokat kapjuk:

$$\begin{aligned} &\frac{dw_1^1(t)}{dt} = \frac{1}{2}Aw_1^1(t), \quad 0 < t \leq \tau, \\ &w_1^1(0) = w_0, \end{aligned} \tag{21}$$

$$\begin{aligned} &\frac{dw_2^1(t)}{dt} = \frac{1}{2}Aw_2^1(t), \quad 0 < t \leq \tau, \\ &w_2^1(0) = w_1^1(\tau). \end{aligned} \tag{22}$$

(Az egyszerűség kedvéért csak az első lépést írtuk le.) Ha az explicit Euler-módszert alkalmazzuk az első (21) feladatra és az implicit Euler-módszert a második (22) feladatra a $\Delta t = \tau$ megválasztással, akkor a következő teljes diszkretizációt nyerjük:

$$\begin{aligned} &\frac{y_1^1 - y_1^0}{\tau} = \frac{1}{2}Ay_1^0; \quad y_1^0 = w_0, \\ &\frac{y_2^1 - y_2^0}{\tau} = \frac{1}{2}Ay_2^1; \quad y_2^0 = y_1^1. \end{aligned}$$

Így a teljes diszkretizációs operátor $C_{\text{tot}} = C_{\text{tot}}(\tau, EE, 1, IE, 1)$ alakja

$$C(\tau) = (I - \frac{\tau}{2}A)^{-1}(I + \frac{\tau}{2}A),$$

amely a jól ismert Crank–Nicolson-módszer.

Megjegyzés. Ez a példa is jól tükrözi, hogy a teljes diszkretizáció rendjének meghatározása bonyolult feladat. Intuitív módon azt gondolnánk, hogy a „leggyengébb láncszem" elve alapján a szeletelés, illetve az alkalmazott numerikus eljárások rendszámának kisebbike hatarozza meg a teljes diszkretizáció rendjét, azaz, ha $rend_{\text{szel}}$ a szeletelés rendje és $rend_{\text{numi}}$ az i-edik szeletelt feladat megoldására alkalmazott numerikus módszer rendje, akkor a teljes diszkretizáció rendje

$$rend_{\text{tot}} = \min\{rend_{\text{szel}}, rend_{\text{num1}}, \ldots rend_{\text{numd}}\}. \tag{23}$$

Ebben az esetben $rend_{\text{szel}} = 0$, mivel a szeletelt operátorok kommutálnak. Továbbá $rend_{\text{num1}} = rend_{\text{num2}} = 1$. Így (23) esetén a teljes rend egy lenne. Ugyanakkor, mint az jól ismert, a Crank–Nicolson-módszer másodrendű.

6.2. Másodrendű Yanenko-módszer

Tekintsük az eredeti (2) Cauchy-feladatot $d = 2$ esetén. A szekvenciális szeletelés és a középponti módszer (trapézszabály) szerinti numerikus integrálással,

valamint a $\tau = \Delta t$ megválasztással ekkor a

$$\begin{aligned} \frac{y_1^{n+1} - y_1^n}{\tau} &= A_1\left(\frac{y_1^{n+1} + y_1^n}{2}\right) \\ y_1^n &= y_2^{n-1} \end{aligned} \tag{24}$$

$$\begin{aligned} \frac{y_2^{n+1} - y_2^n}{\tau} &= A_2\left(\frac{y_2^{n+1} + y_2^n}{2}\right) \\ y_2^n &= y_1^{n+1}, \end{aligned} \tag{25}$$

feladatokat nyerjük, ahol $y_2^{-1} = w_0$ és $n = 0, 1, 2, \ldots, N-1$. Ez az eljárás a másodrendű Yanenko-módszerként ismeretes [19]. Gyakran az A_i operátorok az irányok szerinti felbontásból adódnak. Például amikor A a d-dimenziós Laplace-operátor, akkor A_i az x_i változó szerinti második derivált. Ekkor a fenti algoritmus realizálása egyszerű, mivel lépésenként tridiagonális mátrixú lineáris algebrai egyenletrendszerek megoldását igényli csak. Ezért az irányok szerinti szeletelés elnevezés is használatos.

6.3. Szekvenciális, alternáló Marcsuk-módszer

Jelölje az (24)–(25) Yanenko-módszert $y^{n+1} = \Phi_{A_1 A_2}(y^n)$. A szimmetria megőrzése céljából lépésenként cseréljük meg az A_1 és A_2 operátorok sorrendjét! Ez a

$$y^{n+1} = \Phi_{A_1 A_2}(y^n);\ y^{n+2} = \Phi_{A_2 A_1}(y^{n+1}),\ n = 0, 2, 4 \ldots$$

módszerhez vezet, amelyet az irodalomban szekvenciális, alternáló Marcsuk-módszernek neveznek [14]. Ez tehát megfelel a Strang–Marcsuk-szeletelesű, trapézszabályú numerikus módszert alkalmazó, $\tau = \Delta t$ megválasztású teljes diszkretizáló numerikus módszernek.

6.4. Párhuzamos alternáló módszer

Tekintsük az

$$y^{n+1} = \frac{1}{2}\Phi_{A_1 A_2}(y^n) + \frac{1}{2}\Phi_{A_2 A_1}(y^n)$$

numerikus módszert, amely az irodalomban Swayne-módszerként ismeretes [17]. Könnyen láthatóan ez a módszer megfelel a szimmetrikusan súlyozott szekvenciális szeletelésnek, a középponti numerikus integrálást és $\tau = \Delta t$ lépésközöket választva.

6.5. Lokális egydimenziós sémák

Tekintsük a háromdimenziós hővezetesi egyenletet. Ekkor a Yanenko-módszer a következő alakot ölti:

$$\frac{y_1^{n+1}-y^n}{\tau}=\Lambda_{xx}y_1^{n+1}, \qquad \frac{y_2^{n+1}-y_1^{n+1}}{\tau}=\Lambda_{yy}y_2^{n+1},$$
$$\frac{y_3^{n+1}-y_2^{n+1}}{\tau}=\Lambda_{zz}y_3^{n+1}, \quad y^{n+1}=y_3^{n+1},$$

ahol $n=0,1\ldots$, $y^0=w_0$. Ebben a sémában $\Lambda_{xx},\Lambda_{yy}$ és Λ_{zz} az egydimenziós Laplace-operátor szokásos diszkretizációját jelöli. Ez a módszer megfelel a szekvenciális szeletelésnek, implicit Euler-módszerrel és $\tau=\Delta t$ megválasztással.

7. Összefoglalás

A dolgozatban bevezettük az operátorszeletelés fogalmát és definiáltuk a leggyakrabban alkalmazott módszereket. Elemeztük a különböző tipusú szeleteléseket, mint időbeli diszkretizációs eljárásokat, kitérve azok pontosságára a lokális szeletelési hiba értelmében. Megnéztük, hogy az egyes szeletelt problémákra alkalmazott numerikus eljárások hogyan hatnak ki a teljes diszkretizációra.

A dolgozat terjedelmi okokból nem tér ki a módszer hatékonyságának, illetve a számítógépes realizálásának kérdésére. Itt csak utalunk a légszennyeződési feladatra vonatkozó [3], [6] és [11] dolgozatokra.

Hivatkozások

[1] K. A. Bagrinovskiĭ, S. K. Godunov: *Difference schemes for multidimensional problems*, Dokl. Akad. Nauk SSSR (N.S.) **115**, 431–433 (1957).

[2] M. Bjørhus: *Operator splitting for abstract Cauchy problems*, IMA Journal of Numerical Analysis **18**, 419–443 (1998).

[3] M. Botchev, I. Faragó, Á. Havasi: *Testing weighted splitting schemes on a one-column transport-chemistry model*, Int. J. Environmental Pollution, **22**, 3–16 (2004).

[4] M. Botchev, I. Faragó, R. Horváth: *Application of the operator splitting to the Maxwell equations with the source term*, Appl. Num. Math. (közlésre elfogadva)

[5] P. Csomós, I. Faragó, Á. Havasi: *Weighted sequential splittings and their analysis*, Comp. Math. Appl. **50**, 1017–1031 (2005).

[6] P. Csomós, I. Dimov, I. Faragó, Á. Havasi, Tz. Ostromsky: *Computational complexity of weighted splitting scheme on parallel computers*, International Journal of Parallel, Emergent and istributed Systems, **22**, 137–147 (2007).

[7] K.-J. Engel, R. Nagel: *One-parameter semigroups for linear evolution equations, Graduate Texts in Mathematics*, **194**, Springer, New York (2000).

[8] I. Faragó, Á. Havasi: *The mathemathical background of operator splitting and the effect of non-commutativity*, Lect. Notes Comp. Sci., 2179, Springer Verlag, 264–271 (2002).

[9] I. FARAGÓ, Á. HAVASI: *Consistency analysis of operator splitting methods for C_0-semigroups*, Semigroup Forum, **74**, 125–139, (2007).

[10] I. FARAGÓ, Á. HAVASI: *Relationship between vanishing splitting errors and pairwise commutativity*, Applied Math. Letters, **21** (2008), 10–14.

[11] I. FARAGÓ, K. GEORGIEV, Z. ZLATEVEV: *Parallelization of advection-diffusion-chemistry modules*, Lect. Notes Comp. Sci., 4818, Springer Verlag, Berlin (2007) 28–39.

[12] P. LAX: *Functional Analysis*, Wiley Interscience, (2002).

[13] G. I. MARCHUK: *Some application of splitting-up methods to the solution of mathematical physics problems.* Applik.Mat., **13**, 103–132 (1968).

[14] G. I. MARCHUK: *Splitting and alternating direction methods*, North Holland, Amsterdam, (1990).

[15] G. STRANG: *Accurate partial difference methods I: Linear Cauchy problems*, Archive for Rational Mechanics and Analysis **12**, 392–402 (1963).

[16] G. STRANG: *On the construction and comparison of difference schemes*, SIAM J. Numer. Anal. **5**, 506–517 (1968).

[17] D. A. SWAYNE: *Time dependent boundary and interior forcing in locally one-dimensional schemes*, SIAM Journal on Scientific and Statistical Computing **8**, 755–767 (1987).

[18] J. G. VERWER, W. HUNDSDORFER: *Numerical solution of time-dependent advection-diffusion-reaction equations*, Springer (2003).

[19] N. N. YANENKO: *The method of fractional steps*, Springer, Berlin, (1971).

FARAGÓ ISTVÁN
Eötvös Loránd Tudományegyetem
Matematikai Intézet, Alkalmazott Analízis és Számításmatematikai Tanszék
H-1117 Budapest, Pázmány Péter sétány 1/c
faragois@cs.elte.hu

OPERATOR SPLITTINGS AND THEIR ANALYSIS

ISTVÁN FARAGÓ

Operator splitting is a widely used and successfully applied process, by which a problem of complicated structure can be substituted with a sequence of simpler problems. In this study we present the most important operator splitting methods and touch upon the issues of their computer realisation. We study the numerical behaviour of the procedure and the methods obtained when the split sub-problems are solved numerically. Finally, we set up connections between these combined methods and several other ones known from the literature.

Az Alkalmazott Matematikai Lapok megjelenését támogatja
a Magyar Tudományos Akadémia Könyv- és Folyóiratkiadó Bizottsága.

A kiadásért felelős a BJMT főtitkára
Szedte és tördelte Éliás Mariann

Nyomta a Nagy és Társa Kft., Budapest
Felelős vezető: Fódi Gábor

Budapest, 2009
Megjelent 18 (A/5) ív terjedelemben
250 példányban
HU ISSN 0133-3399

ÚTMUTATÁS A SZERZŐKNEK

Az Alkalmazott Matematikai Lapok csak magyar nyelvű dolgozatokat közöl. A közlésre szánt dolgozatokat e-mailen az aml@math.elte.hu címre kérjük elküldeni az ábrákat tartalmazó fájlokkal együtt. Előnyben részesülnek a LaTeX-ben elkészített dolgozatok.

A kéziratok szerkezeti felépítésének a következő követelményeket kell kielégíteni. A fejlécnek tartalmaznia kell a dolgozat címét és a szerző teljes nevét. A fejléc után egy, képletet nem tartalmazó, legfeljebb 200 szóból álló kivonatot kell minden esetben megadni. A dolgozatot címmel ellátott szakaszokra kell bontani, és az egyes szakaszokat arab sorszámozással kell ellátni. Az esetleges bevezetésnek mindig az első szakaszt kell alkotnia. Az irodalomjegyzék után, a kézirat befejezéseképpen fel kell tüntetni a szerző teljes nevét és a munkahelye (illetve lakása) pontos címét. A dolgozatban előforduló képleteket szakaszonként újrakezdődően, a képlet előtt két zárójel közé írt kettős számozással kell azonosítani. Természetesen nem szükséges minden képletet számozással ellátni. Az esetleges definíciókat és tételeket (segédtételeket és lemmákat) ugyancsak szakaszonként újrakezdődő, kettős számozással kell ellátni. Kérjük a szerzőket, hogy ezeket, valamint a tételek bizonyítását a szövegben kellő módon emeljék ki. Minden dolgozathoz csatolni kell egy angol, német francia vagy orosz nyelvű, külön oldalra gépelt összefoglalót.

Mind az ábrákat, mind a lábjegyzeteket a dolgozat szakaszokra bontásától független, folytatólagos arab sorszámozással kel ellátni. Az ábrák elhelyezését a dolgozat megfelelő helyén, széljegyzetként feltüntetett, ábraazonosító sorszámokkal kell megadni. A lábjegyzetekre a dolgozaton belül az azonosító sorszám felső indexkénti használatával lehet hivatkozni.

Az irodalmi hivatkozások formája a következő. Minden hivatkozást fel kell sorolni a dolgozat végén található irodalomjegyzékben, a szerzők, illetve a társszerzők esetén az első szerző neve szerint alfabetikus sorrendben úgy, hogy a cirill betűs szerzők nevét a Mathematical Reviews átírási szabályai szerint latin betűsre kell átírni. A folyóiratban megjelent cikkekre [1], a könyvekre [5], a kötetben megjelent dolgozatokra [4], a disszertációkra [3] és a gépi program leírásokra [2] a következő minta szerint kell hivatkozni:

[1] Farkas, J.: *Über die Theorie der einfachen Ungleichungen.* Journal für die reine und angewandte Mathematik **124**, (1902) 1–27.

[2] Kéri, G.: „DUALSIMP", rutin a CDC 3300-ás gépekre (Magyar Tudományos Akadémia Számítástechnikai és Automatizálási Kutató Intézete, CDC 3300 felhasználói ismertetők 2. 1973. május) 19–20.

[3] Prékopa, A.: *„Sztochasztikus rendszerek optimalizálási problémáiról"*, doktori értekezés. Magyar Tudományos Akadémia, Budapest, 1970.

[4] Prabhu, N. U.: *„Recent research on the ruin problem of collective risk theory"*, in: Inventory Control and Water Storage. Ed. A. Prékopa (János Bolyai Mathematical Society and North-Holland Publishing Company, Amsterdam–London, (1973) 221–228.

[5] Zoutendijk, G.: *Methods of Feasible Directions* (Elsevier Publishing Company, Amsterdam and New York, 1960).

A dolgozatok szövegében az irodalmi hivatkozás számait szögletes zárójelben kell megadni, mint például [5] vagy [4, 76–78]. A szerzők a dolgozatukról 50 darab ingyenes különlenyomatot kapnak. A dolgozatok után szerzői díjat az Alkalmazott Matematikai Lapok nem fizet.

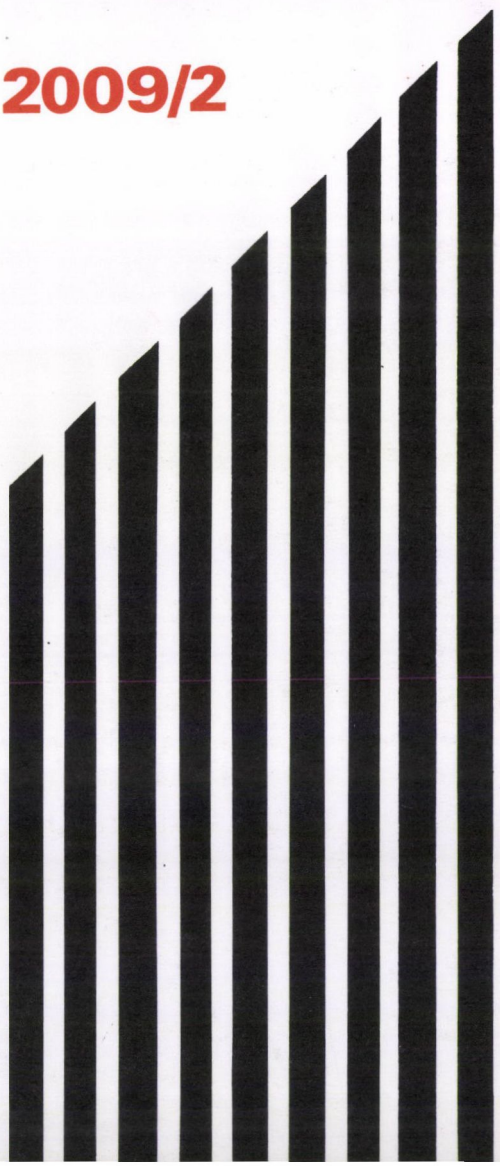
TARTALOMJEGYZÉK

INDEX

Alkalmazott matematikai lapok

2009/2

A MAGYAR TUDOMÁNYOS AKADÉMIA
MATEMATIKAI TUDOMÁNYOK
OSZTÁLYÁNAK KÖZLEMÉNYEI



26.

KÖTET

ALKALMAZOTT MATEMATIKAI LAPOK

A MAGYAR TUDOMÁNYOS AKADÉMIA MATEMATIKAI TUDOMÁNYOK OSZTÁLYÁNAK KÖZLEMÉNYEI

26. kötet
Szerkesztőség és kiadóhivatal: 1027 Budapest, Fő u. 68.

Az Alkalmazott Matematikai Lapok változó terjedelmű füzetekben jelenik meg, és olyan eredeti tudományos cikkeket publikál, amelyek a gyakorlatban, vagy más tudományokban közvetlenül felhasználható új matematikai eredményt tartalmaznak, illetve már ismert, de színvonalas matematikai apparátus újszerű és jelentős alkalmazását mutatják be. A folyóirat közöl cikk formájában megírt, új tudományos eredménynek számító programokat, és olyan, külföldi folyóiratban már publikált dolgozatokat, amelyek magyar nyelven történő megjelentetése elősegítheti az elért eredmények minél előbbi, széles körű hazai felhasználását. A szerkesztőbizottság bizonyos időnként lehetővé kívánja tenni, hogy a legjobb cikkek nemzetközi folyóiratok különszámaként angol nyelven is megjelenhessenek.

A folyóirat feladata a Magyar Tudományos Akadémia III. (Matematikai) Osztályának munkájára vonatkozó közlemények, könyvismertetések stb. publikálása is.

A kéziratok a főszerkesztőhöz, vagy a szerkesztőbizottság bármely tagjához beküldhetők. A főszerkesztő címe:

Páles Zsolt, főszerkesztő
1027 Budapest, Fő u. 68.
A folyóirat e-mail címe: aml@math.elte.hu

Közlésre el nem fogadott kéziratokat a szerkesztőség lehetőleg visszajuttat a szerzőhöz, de a beküldött kéziratok megőrzéséért vagy továbbításáért felelősséget nem vállal.

Az Alkalmazott Matematikai Lapok előfizetési ára évfolyamonként 1200 forint. Megrendelések a szerkesztőség címén lehetségesek.

A Magyar Tudományos Akadémia III. (Matematikai) Osztálya a következő idegen nyelvű folyóiratokat adja ki:

1. Acta Mathematica Hungarica,
2. Studia Scientiarum Mathematicarum Hungarica.

Alkalmazott Matematikai Lapok **26** (2009), 273-287.

MODELLVEZÉRELT SZOFTVEREK KÉSZÍTÉSE II.

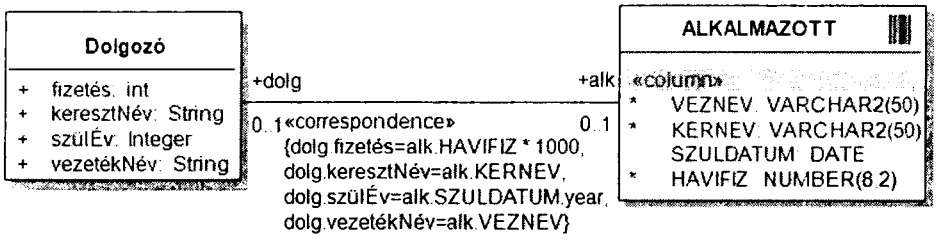
KILIÁN IMRE

1. Bevezetés

A cikk első részében (Alkalmazott Matematikai Lapok, **26** (2009), 1–7. oldal) bemutattuk a modellvezérelt szoftverek általános felépítését, és elemeztük azok alkalmazási lehetőségeit. Ilyenek a szoftver fejlesztési és tervezési gyakorlatban többféle helyzetben is felmerülnek, ezek kielégítése kétszintű szerkezettel – a modellszint tárolásával könnyebb és kevésbé kockázatos lehet. A cikk jelen, második szakaszában erre vonatkozó példákat mutatunk be.

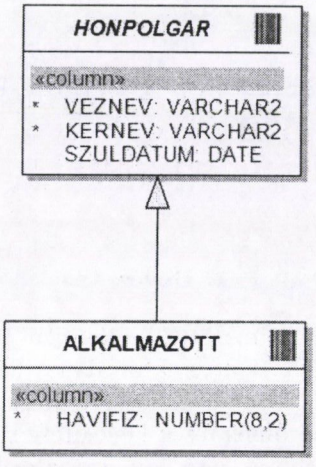
1.1. UML testreszabása

A példákban az UML következő olyan testreszabásait alkalmazzuk, amelyek nem magától értetődőek.



1. ábra. UML osztályok megfeleltetése

Megfeleltetés (correspondence): A megfeleltetés olyan sztereotípus, amely két osztály közötti 0..1–0..1, vagy szűkebb többszörösségű kapcsolatra alkalmazható, és az egymásnak megfeleltethetőség intuitív fogalmát fejezi ki, vagyis azt, hogy mindkét osztály példányainak van egy olyan részhalmaza, amelyek egymásnak kölcsönösen egyértelműen megfeleltethetők. Ha ez a részhalmaz mindkét esetben a teljes példányhalmaz, akkor a többszörösség éppen 1–1, a megfeleltetés pedig *szoros* lesz. Ha 0..1–1, akkor a bal kapcsolatvég minden egyes példányához létezik megfelelő elem a jobb kapcsolatvégről. Ha 1–0..1, akkor fordítva. Ha a többszörösség mindkét kapcsolatvégen 0..1, akkor a megfeleltetés *laza*, mindkét kapcsolatvégen lehetnek olyan példányok, amelyek nem vesznek részt a megfeleltetésben, vagyis amelyeknek nincs megfelelőjük a másik oldalon.



2. ábra. Absztrakt relációs táblák

A megfeleltetést a bemutatott példákban elsősorban egyes háttérben tárolt (perzisztens) UML osztályok és az őket megvalósító relációs táblák összekapcsolására használjuk. A megfeleltetés egyéb megkötéseit a kapcsolatra írt OCL megszorítással adhatjuk meg.

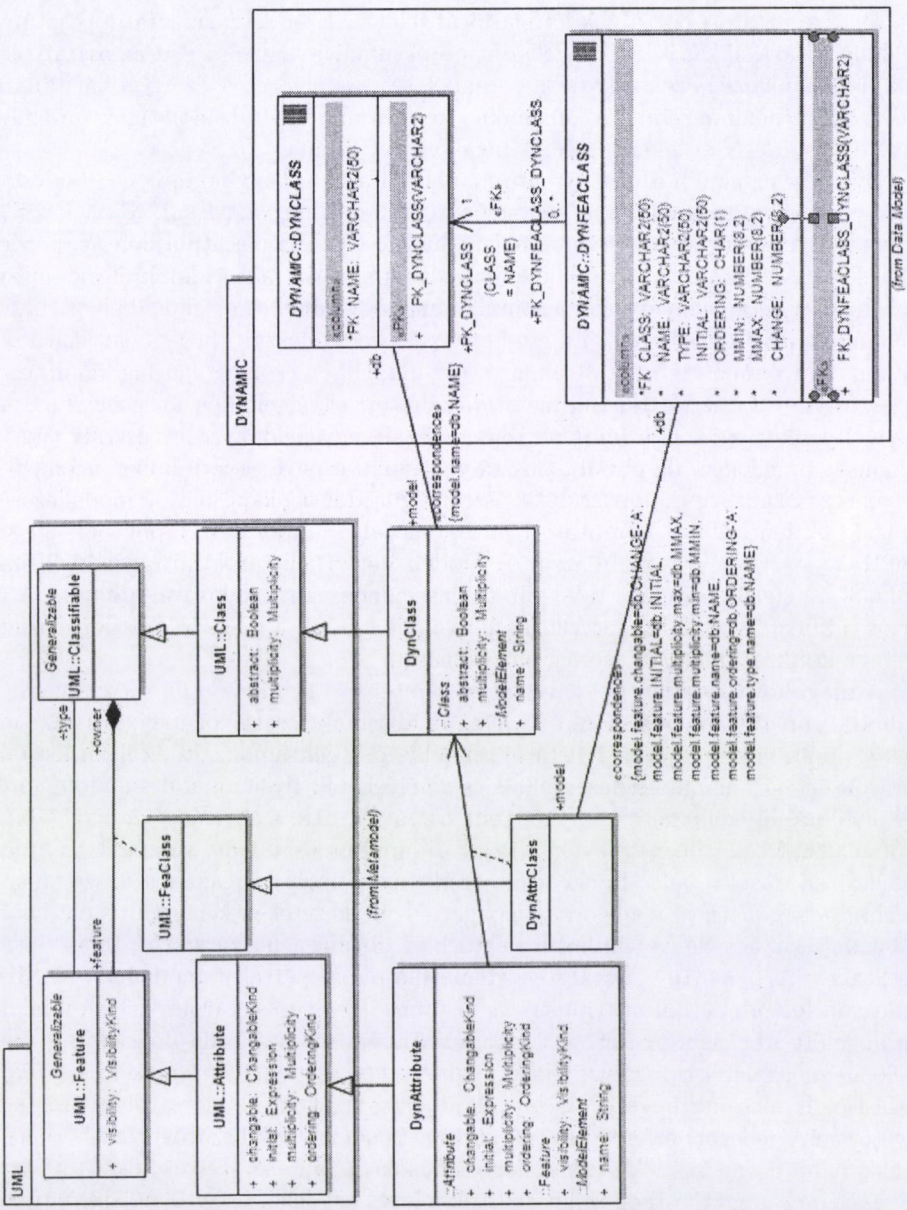
Öröklődés relációs táblák között, és absztrakt relációs táblák: Az egyes relációs táblákban ismétlődően megvalósítandó oszlopszekvenciákat absztrakt relációs táblákkal írjuk le. Azokat a konkrét táblákat, amelyekben az absztrakt táblák oszlopait szeretnénk látni, az UML öröklődés kapcsolóval leszármaztatjuk az absztrakt ős-táblából.

2. Kétszintű rendszerek: modellek dinamikus kezelése

A kétszintű működés esetében tehát a modellek futásidejű létrehozása, módosítása, esetleg törlése szükséges. Az alapprobléma egyes részeit a következő módon oldhatjuk meg.

2.1. Dinamikus tulajdonságok

A leghétköznapibb ilyen igény az, amikor az objektumközpontú felépítmény megengedte rögzített tulajdonságlista nem elegendő, mert a felhasználó megköveteli, hogy egyes objektumok tulajdonságlistája futásidőben módosuljon, ill. kibővülhessen. Az ilyen igények egyedi megoldásokon alapuló kielégítése helyett a modell részleges tárolása alapján az alább vázolt egységes és hatékonyabb megoldás javasolható.



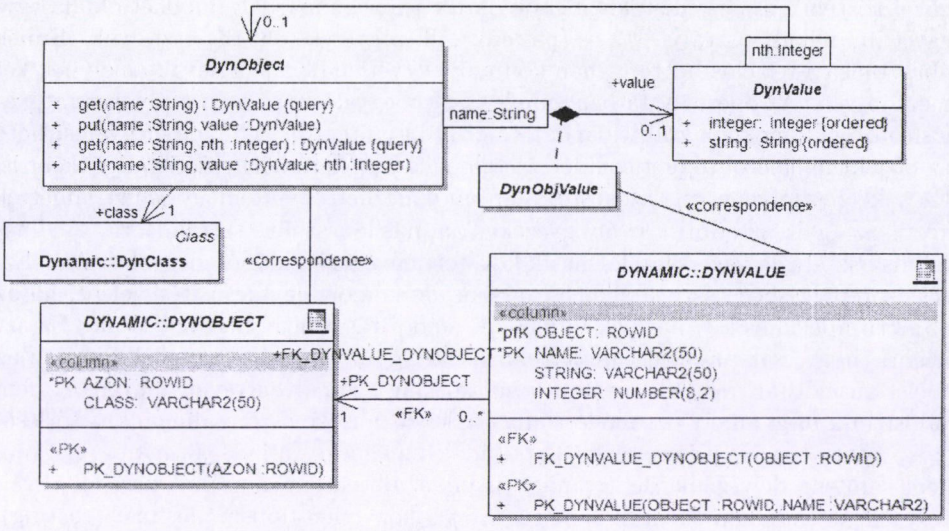
3. ábra. Dinamikus tulajdonságlista – modellszint

A megoldás a jelen esetben is kétszintű: a modell tulajdonságleíró részletének tárolása a metamodell egy részének beprogramozását, ill. a metamodell kiterjesztését, leszármaztatott osztályok létrehozását jelenti. Ezen leszármaztatott osztályok példányosításával kapjuk a modellszint objektumait, vagyis azokat az osztályokat, tulajdonságokat és ezek egymáshoz rendelését, amelyeket a szoftverben dinamikusan szeretnénk kezelni. A dinamikus kezelés miatt a tulajdonságok futásidejű felvételét, törlését és lekérdezését is meg kell valósítani.

Ilyenkor az alábbi ábrán látható modellből a `DynAttribute`, `DynClass` osztályokat és a `DynFeatureClass` kapcsolóosztályt kell megvalósítani. Ezek közül az előbbi kettő esetleg el is hagyható akkor, ha a `DynClass` metatulajdonságai érdektelenek, vagyis ha nem akarunk különbséget tenni absztrakt és konkrét metaosztályok között, netán nem érdekel bennünket az, hogy hányszorosan példányosítható a dinamikusan kezelt osztály. Ez esetben azt is feltételezzük, hogy nem akarunk a dinamikus tulajdonságlistát alkalmazó rész-modellben csomagokat használni, ezért az osztálynév (a `name:String` metatulajdonság) egyértelműen azonosítja az osztályt. Ilyenkor ugyanis a `DynAttribute` osztály egy példánya egy osztály egy tulajdonságát jellemzi, de ebből az osztályt csupán a neve egyértelműen azonosítja. A `DynAttrClass` kapcsolóosztály az osztály-tulajdonság kapcsolatok modellezésére szolgál. A fenti ábrán bemutatott megoldás teljes abból a szempontból, hogy a `DynClass` osztályt is tartalmazza. A modell háttérbeli tárolására a modellosztályoknak megfelelő relációs táblák, ill. táblaszegmensek megadásával utalunk. A relációs táblákat absztraktként vettük fel, vagyis csupán a megadott oszlopkészletet kell egy konkrét táblában esetleg megvalósítani.

A megoldást az adat/példányszinten szintén elő kell készíteni. Erre a célra az alábbi ábrán látható `DynObject-DynValue` absztrakt osztályokat biztosítjuk, melyek a dinamikus tulajdonságlistával rendelkező objektumok, ill. az ilyen listában ábrázolt értékek alapműködését adják, és amelyeket az ilyen működésmódot igénylő osztályoknál örökölni kell. A `DynObject` osztály a `class` egyirányban navigálható kapcsolaton keresztül megnevezi a saját dinamikus osztályát, valamint tulajdonságelérő eljárásokat ad. Ezekre tulajdonképpen nincs is igazán szükség, hiszen a tulajdonság-értékek a megadott navigációs szerkezeteken keresztül is elérhetők. A `DynObject` egy névvel indexelt vektorban tárolja a hozzátartozó tulajdonság-értékeket. A `DynValue` osztály egyetlen tulajdonság-értéket modellez. Ez diszjunktívan felépített (union) típusú, vagyis bármilyen típust tárolni kell tudnia. Az ábrán jelölt `string:String`, ill. `integer:Integer` osztálytulajdonságok a megfelelő nevű skaláris típusokat jelentik, amelyeket esetleg bővíteni kell, ha újabb skalárisokat akarunk bevezetni. A tulajdonság-értékek esetleges többszörösségét (gyűjtemény jellegét) a fenti két tulajdonság rendezett többszörösségével érhetjük el akkor, ha a fent modellezett skaláris típusokról van szó. Egyéb, összetett értékek esetében az `nth:Integer` kiválasztókifejezés azt jelzi, hogy `DynObject` típusú adatok egy vektoráról van szó.

Az ábrázolás bonyolultságát az „union" típus, valamint a tulajdonságértékek többszörösségének megvalósítása adja. Ha az adott alkalmazásban a többszörösséget egyszeres (vagy üres) értékre (0..1) korlátozzuk, akkor az „nth" indexelés meg-



4. ábra. Dinamikus tulajdonságlista – példányok szintje

szűnhet mind a `DynValue-DynObject` kapcsolat, mind a `DynValue` osztály többszörös tulajdonságainak tekintetében, Ha viszont az értékek típusait pl. stringgé egyértelműen konvertálható alaptípusokban rögzítjük, akkor az egész `DynValue` osztály egyetlen `String` értékre egyszerűsödhet.

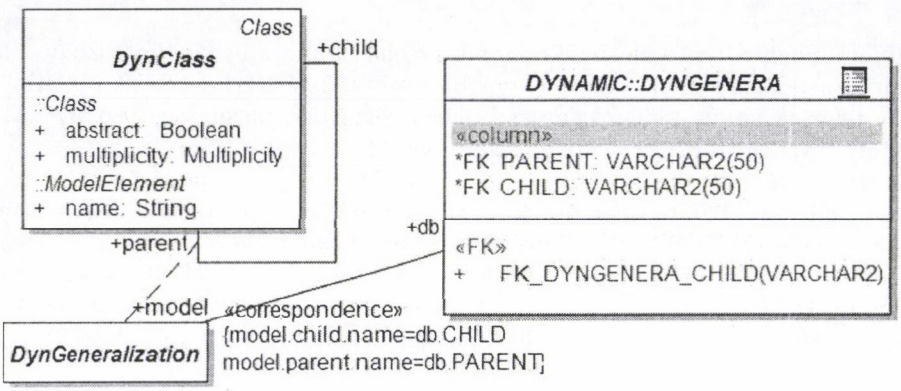
Ha egy osztály dinamikus tulajdonságkészletének megvalósításáról, és az objektum adatbázisban történő tárolásáról van szó, akkor a `DYNOBJECT` absztrakt táblától történő öröklődés figyelembevétele a háttértárban mindössze az osztálynévnek megfelelő új `CLASS` mező felvételét jelenti, hiszen a másik, az `AZON` azonosítómező valószínűleg amúgy is létezik. Ezen túl valamint a `DynObject-DynValue` kapcsolatot kell egy újonnan létrehozott táblával modellezni (az ábrán: `DYNVALUE`).

2.2. Dinamikus osztályszerkezet

A dinamikus osztályszerkezet megvalósítása esetén az egyik megoldás szerint a létrejövő dinamikus osztályok példányai mind egyetlen rögzített őstípushoz tartoznak, és a dinamikus osztályok futásidőben csupán egy osztálykötődést jelző tulajdonsággal, dinamikusan változó tulajdonságszerkezettel, ill. dinamikusan létrejövő függvénykészlettel vannak megvalósítva. Ezt az ősosztályt az adott programnyelven meg kell valósítani, a többi, dinamikusan létrehozott osztály, valamint a tulajdonságkészletük is már csak képzetes módon, futási időben áll elő, csak dinamikusan hozható létre.

A dinamikus osztályszerkezet megvalósítása a modellszinten a metamodell hasonló nevű kapcsolatát kiterjesztő `DynGeneralization` kapcsolóosztály segítsé-

gével történik. Ez egy kétoldalú kapcsolatot jelent, amelynek mindkét oldala egy-egy (dinamikus) osztály – az ős (`parent`), ill. a gyerek (`child`) osztályok. Ennek ábrázolása a relációs háttértárban könnyű – egyetlen olyan táblával, amelynek két mezője van. Az operatív tárban ennek egyik megvalósítása az osztályokat megvalósító objektumokhoz kötött `parent` és `child` gyűjteményekkel, de megvalósítható az objektumhálótól függetlenül létező, globális hatókörű környezeti változóként is. Ez a két osztálynév, ill. azonosító alapján címezhető gyűjtemény lenne, amelyek közül az egyik az adott osztály gyerekeit, a másik a szüleit tartalmazza. A dinamikus osztályszerkezet-kezelés másik követelménye az öröklődés megvalósítása. Ezzel összefüggésben ésszerű alkunak látszik, de komoly egyszerűsítést jelent, ha az objektumpéldányok típusának futásidejű módosítását nem engedélyezzük. Ez azt jelenti, hogy csak az objektum létrehozásakor gyűjtjük össze a modell alapján az tulajdonságlistát, de példányokra vonatkozólag semmilyen olyan műveletet nem valósítunk meg, amely ezt módosítaná. A leszármazási viszony dinamikus törlését – az adott típusú objektumok állapotának meghatározatlansága miatt – egyébként sem engedélyezzük, de ugyanígy tiltjuk az objektum típusának futásidejű változtatását – a konkretizálást éppúgy, mint a közvetlen típusmódosítást (casting). Tehát a választott modellben a tulajdonság mindig az első hivatkozásakor, az ak-



5. ábra. Dinamikus osztályszerkezet megvalósítása

kori osztályszerkezet alapján, a korai típuslekötés elveinek megfelelően jön létre, ami későbbiekben nem változik. Első közelítésben feltételezzük azt is, hogy az eredeti tulajdonságkészletet újabb öröklési utak felvétele sem változtatja.

A leírt egyszerűsítések lényegesen megkönnyítik a vonatkozó megvalósítást, és feltehetően a szóba jövő szoftverek túlnyomó részénél elegendőek lehetnek. Mindazonáltal az egész csupán megvalósítási könnyebbség, vagyis a fent vázolt adatszerkezet alapján a tényleges megvalósítás nem jelenthet elvi akadályt.

2.3. Dinamikus kapcsolatok és osztályfüggvények

A kapcsolatok és az osztályfüggvények dinamikus kezelését konkrét terv szintjén nem gondoltuk végig. Ezért ebben a szakaszban csupán az ezzel kapcsolatos alapvető meggondolásokat szeretnénk bemutatni.

Modellünk kapcsolatainak kezelése – csupán a megvalósítás szempontjából – felfogható az osztálytulajdonságokhoz hasonlóan is. Vagyis ha a modellszinten a kapcsolatok kezelése is követelmény, a legkézenfekvőbb azokat tulajdonságokká – összetett típusú objektumértékű tulajdonságokká átalakítva megvalósítani.

Osztályfüggvények kezelése lényegesen komolyabb gondokat vet fel. A futásidőben történő létrehozhatóság követelménye miatt egy futásidőben beprogramozható és beszerkeszthető megoldásra van szükség, ami behatárolja a megvalósítás nyelvét (pl. Java, Python, Prolog alkalmas ilyesmire). Olyan objektumközpontú nyelv esetén, amely ilyesmit nem támogat, a dinamikus osztályszerkezeten keresztül újabb függvények futásidejű hozzászerkesztése nem megoldható.

2.4. A modellszint és az alkalmazói szint összefüggése

Ezen a ponton egy pillanatra álljunk meg, hogy megkötést tegyünk a futásidőben a metamodell példányaként változó modell és a változó modell példányaként változó objektumháló összefüggésére, arra, hogy a modell módosítása milyen kihatással van az adatszintre. Erre a következő lehetőségeink kínálkoznak:

- A modellre vonatkozóan minden változtatást a program betöltésének és futtatásának ideje között egy *modellrögzítési fázisban* teszünk meg. Vagyis bár a modell egy része az alkalmazói programban van tárolva, de mégsem tekinthető dinamikusan változónak, hiszen egy programindítás utáni betöltési, ill. esetleges módosítási fázis utáni állapot befagy, és tovább már nem módosítható.
- A modellre vonatkozóan a *futásidőben is megengedjük a bővítést*, vagyis újabb modellobjektumok felvételét, *de a törlésüket nem.* Ez megfelel az ún. *monoton logikai* megközelítésnek, amikor is a kezelt tudásanyag csak nőhet, de sosem csökken. Gyakorlati szempontból ezzel a megkötéssel megtakaríthatjuk a modellelemek törléséből fakadó különbözőféle következményhatások (pl. törölt típusú objektumpéldány) figyelembevételét és megvalósítását.
- *Bármilyen modellműveletet megengedünk,* a nem monoton törlési művelet összes tovagöngyölődő hatását figyelembe vesszük, és megvalósítjuk (pl. tulajdonság törlésekor az összes hivatkozó tulajdonságpéldányt töröljük). Ez általánosságban véve az összes törölhető elemre (`DynAttribute`, ill. `DynClass`) mutató, csak egy irányban navigálható hivatkozási kapcsolat kétirányúvá tételét jelenti, amelyen keresztül a hivatkozó objektumokat a hivatkozás megszűntéről értesíteni lehet.

Az első változat melletti döntés valószínűleg túlságosan erős megszorítás. A modell- és az alkalmazói szint közötti kapcsolat túlságosan befagyott, merev, egy ilyen kapcsolat nem is igazán indokolja a modell- és az adatszint együttes tárolását és kezelését. A harmadik változat viszont túlságosan bonyolultnak, áttekinthetetlennek tűnhet, különösen akkor, ha – mint a következő szakaszban olvasható lesz – nemcsak az osztály-tulajdonság kapcsolatot, hanem ennél több modellelemet is dinamikusan kívánunk kezelni.

A két véglet között jó alkunak tűnik a középső változat, különösen úgy, hogy ebből kiindulva bizonyos megszorításokkal a harmadik változat is megvalósítható. Eszerint a törölt adatot nem távolítjuk el, csupán egy töröltséget jelző tulajdonságot állítunk át bennük. A törlésről a hivatkozó objektumokat nem értesítjük szinkron módon, hanem csak késleltetve: minden műveletnél ellenőrizzük a töröltségi állapotot is, és amennyiben ezt a műveletek beállítva találják, akkor egyrészt ezt az állapotot átveszik, másrészt elvégzik a szükséges hibakezelési vagy egyéb műveleteket.

2.5. Modellvezérelt lekérdezés

A modellszintről vezérelt adatcsere legtipikusabb/legáltalánosabb megvalósítása egy objektum-orientált lekérdezési nyelv alkalmazása. Ezek – objektumközpontú programnyelvekhez hasonlóan – igen hasonlóak egymáshoz, lényegében mindegyik az ODMG OQL nyelv valamilyen változatának tekinthető [4]. Az OQL maga – nyitott magú nyelv, csak a legalapvetőbb nyelvtani szerkezetekkel, amely a hagyományos SQL-hez hasonlóan a következő résznyelveket tartalmazza:

- *Lekérdező nyelv*: A résznyelv valójában alapinformációkból – osztálypéldányokból – történő adat-átalakítás absztrakt leírására alkalmas, és a használatát tekintve kétféleképpen hajtható végre:
  - *Hagyományos, szinkron értelemben:* Az információigény fellépésének időpontjában a lekérdezés kiértékelésre, végrehajtásra kerül úgy, hogy a programvégrehajtás megvárja a lekérdezés kiértékelését.
  - *Aszinkron értelemben*, amikor az alapadatok változásához kapcsolt triggerrel indítjuk a lekérdezés kiértékelését, amely a programhoz magához szintén aszinkron – eseményvezérelt módon érkezik.
- *Módosító nyelv*: A résznyelv a háttérben tárolt adatok módosítására – törlésére, cseréjére és létrehozására alkalmas.

2.6. A kétszintű/dinamikus működés további eszközei

A dinamikus működés megoldásához néhány további eszköz szükséges.

Mivel a kezdetben ismert osztály- és tulajdonságszerkezet dinamikus, vagyis nem a tervezés-fordítás ciklus folyamán, hanem a program futása során jön létre, ill. a háttérbeli tárolás miatt esetleg a program indulásakor betöltődik, ezért esz-

közt kell adnunk a modell kezdeti betöltésére (populációjára), valamint a modell futásidejű bővítésére is.

Az *adatbázis kezdeti feltöltése* esetleg egy megfelelően előkészített adatbázis importtal helyettesíthető akkor, ha a kezdeti modell kicsi és egyszerű. A dinamikus módosíthatóság azonban csak úgy biztosítható, ha erre valamiféle eszközt adunk. Ha csak igen kismérvű módosítások képzelhetők el, akkor valami egyszerű grafikus felület is elegendő lehet. Ha a módosítási igényeket nem lehet igen-igen egyszerű mértékre korlátozni, akkor célszerű egy külső *UML részmodellt leíró programnyelv*, ill. a hozzá tartozó *formátumátalakító szoftver*_eszközök (kigeneráló, ill. elemző) használata célszerű, amely egyben a kezdeti modellmegadás feladatát is megoldja. Ehhez mintául szolgálhat a SILK projekt SILAN nyelve [6]. Egy ilyen nyelv általános megadását olvashatjuk az OMG kibocsátott Human-Usable Textual Notation [5] szabványában is. CASE eszközzel (pl. Rational Rose) történő modellcseréhez az XML alapú XMI modell-leíró nyelv használata célszerű.

3. Kétszintű szoftverfelépítmény természetvédelmi alkalmazásokban

Természetvédelmi tárgyú alkalmazói programok térképi adatokat tárolnak, amelyekhez a legkülönbözőbbféle szöveges, ill. számszerű adatokat rendelnek: elsősorban a terepi munkát végző kutatók által végzett észlelések ezek az adatok, amelyekhez dokumentumfelvételeket (álló és mozgó fényképeket, hangfelvételt stb.) is kapcsolhatnak.

A természetvédelemben jobbára a legtöbb üzleti alkalmazásban használatos feladatok merülnek fel. Ami mindezeken túlmutat, az egyrészt a földrajzi adatok kezelése, másrészt pedig az élőlények Linné-féle rendszertanának tárolására vonatkozó igény. Az előbbi a földrajzi rendszerek kialakítási stratégiájának megfelelően részben modellvezérelt módon történik, de mindezek megoldása készen van, vagyis legfeljebb testreszabási, ill. modellkészítési és alkalmazásgenerálási feladataink maradnak.

A Linné-féle rendszertan taxonszerkezetére viszont kiválóan illeszkednek a dinamikus tárolás követelményei. Emellett azonban egy további kikötés is van. Mivel a biológus kutatók fáradhatatlan munkája eredményeképpen a taxonszerkezet változik – fajok összeolvadnak, szétválnak, esetleg újakat fedeznek fel, a rendszerben az *időfüggés* megvalósítása is szükséges. A korábban rögzített adatok értelmezéséhez ezért meg kell oldani azt, hogy a rendszert bármilyen múltbeli időpontra vonatkoztatva is működtetni tudjuk.

3.1. Időfüggés/időgép

Az időgép egy olyan működést jelent, amelyben a rendszer állapotát múltbéli időkre visszamenőleg is le lehet kérdezni. Az időgép általánosságban többféle módon is megvalósítható attól függően, hogy az időfüggés az adatok milyen körére

terjed ki, és milyen finomság a követelmény. Az alábbi leírást tekinthetjük egy programtervezési mintának is: annyiban mégis csak „elő-minta" (proto-pattern), hogy széles alkalmazási körről (legalább 3 sikeres és működő alkalmazás) nem számolhatunk be.

Az időfüggés finomságától függően az alábbi kétféle időfüggést lehetséges megkülönböztetni:

- ha az időfüggő adatok folytonosak, de legalábbis az értékváltozás nagyon sűrű, akkor *folytonos időfüggésről* beszélünk;
- ha az adatok értékei bizonyos események hatására változnak meg, ami után bizonyos ideig (a következő esemény bekövetkeztéig) az értékük állandó marad, akkor *diszkrét időfüggésről* beszélünk, ilyen megoldást tárgyal a jelen tanulmány is.

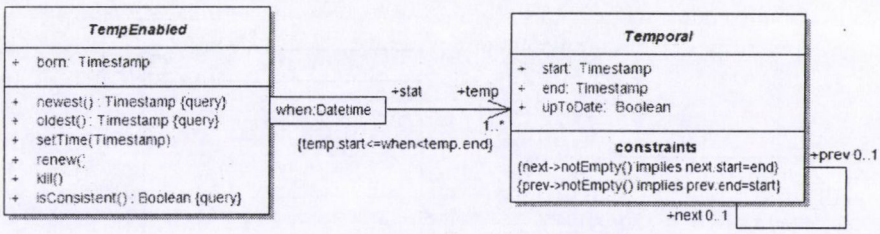
Az időfüggés ritkán terjed ki az adatok teljességére. Attól függően, hogy mely adattartományok időfüggőek, az alábbi megoldások képzelhetők el:

- ha az időfüggés mégiscsak közel teljes körű (minden változik), és sok érték változik egyidejűleg, de nem túl gyakran, akkor valószínűleg a teljes adatállomány történeti mentése a célszerű, vagyis visszamenőlegesen több generációs teljes adattár-állományokat érdemes mentenünk, ezt *állomány szintű időfüggésnek* nevezzük, egy ilyen megoldás a konfiguráció- és verziókezelő rendszerekéhez hasonló működést jelent;
- ha az időfüggés egyes objektumokra, azok összes, de legalábbis több tulajdonságára vonatkozik, ilyenkor *objektum szintű időfüggésről* beszélünk;
- ha az időfüggés csupán egyes objektumok egyes tulajdonságaira vonatkozik, akkor *tulajdonság szintű időfüggésről* beszélünk.

Az alábbiakban az objektum szintű adatfüggés megvalósítására mutatunk be javaslatot. Ebből a tulajdonság szintű adatfüggés már könnyen származtatható, a fenti megjegyzések miatt az állomány szintű adatfüggés viszont más jellegű megoldást igényel.

3.2. Objektum szintű időfüggés

Az objektum szintű időfüggés megvalósítására a fent bemutatott absztrakt osztályokat vezetjük be. Az ilyen objektumok képzeletben két részre bonthatók: a teljesen időfüggetlen jellemzőket (`TempEnabled`) és a teljesen időfüggőket (`Temporal`) tartalmazóra. A kettő egymással a `stat-temp` kapcsolaton keresztül kommunikál, amelynek az időfüggő (`temp`) oldala 1-nél nem kisebb többszörösségű. Ez egy gyűjtemény, amely a diszkréten időfüggő adatokat tartalmazza, és a `when:Datetime` kiválasztókifejezésen keresztül indexelhető. A gyűjtemény elemei rögzítik a saját érvényességi tartományukat (`start-end`), amelyek között átfedés nem lehet, és a következő intervallum kezdete mindig megegyezik a megelőző intervallum végével (ld. a `Temporal` osztály megszorítását). A gyűjtemény indexelésénél azt az idő-



6. ábra. Időgép megvalósítása

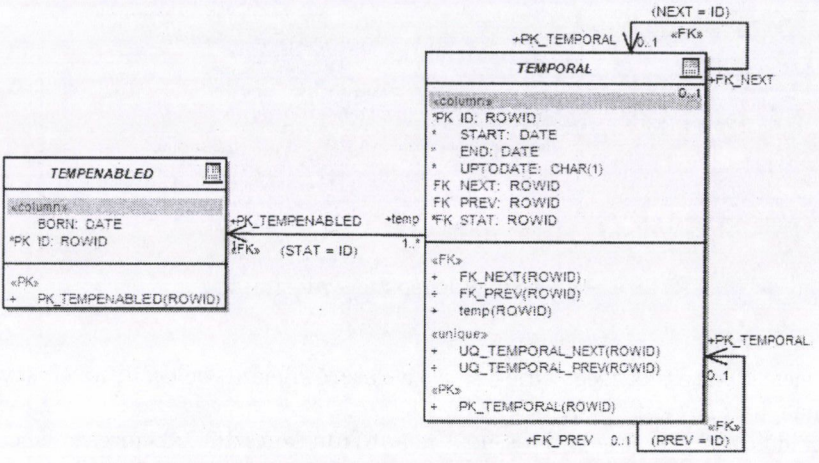
függő elemet keressük meg, amelyre a kiválasztókifejezés beleesik az érvényességi tartományba (`{start<=when<end}`).

Az időfüggő részek a prev és next mutatókon keresztül kétszeresen láncolt listába szervezhetők: ez nem tartozik szigorúan a követelményekhez, de lényegesen megkönnyítheti pl. történelmi összefoglalók készítését.

A `TempEnabled` osztály tartalmazza az objektum időfüggetlen törzsét és az időfüggés kezelésére alkalmas eljárásokat:

- Időlekérdező eljárások (`newest()`, `oldest()`). Ezek a megadott időpontok lekérdezésére alkalmasak. Az adott időpontbeli értékek a kiválasztókifejezés szerinti indexeléssel érhetők el (OCL2).
- referenciaidőpont beállítása (`setTime()`). A rendszer alapértelmezés szerint az aktuális időpontban működik. Az időpont-beállító eljárást akkor használjuk, ha a rendszert mégis visszamenőleges üzemkódban kívánjuk működtetni.
- Az objektum frissítése (`renew()`). Az egyes tulajdonság-értékek változtatása egyöntetű módon történik mind az időfüggetlen, mind az időfüggő objektumszegmens esetén. Amennyiben az időfüggő tulajdonságokat átállítanánk, az `upToDate` tulajdonság önműködően `False` értéket vesz fel. A frissítés eljárás meghívásával az objektumnak egy újabb időfüggő példányát (a `Temporal` gyűjteménynek újabb elemét) hozzuk létre, és az `upToDate` bitet egyidejüleg `True` értékre állítjuk.
- Az objektum élettörténetének lezárása (`kill()`).
- (`isConsistent()`) Az objektumra és a belőle navigációval elérhető társobjektumokra időbeli konzisztencia-ellenőrzést hívunk meg. A megszüntető művelet következtében ugyanis előfordulhat, hogy a társobjektumok az adott (aktuális vagy a beállított) időpontban nem léteznek már, vagy nem léteztek még.

Az említett két osztály mindegyike absztrakt, vagyis önmagukban nem példányosíthatók. Konkrét esetben az időfüggést megkövetelő osztályokat a fenti osztályokból származtatjuk le.



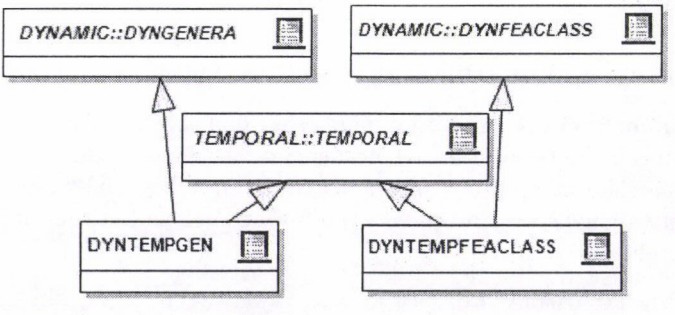
7. ábra. Időgép leképezése relációs adatbázisba

Az objektum szintű időfüggést az osztálydiagramnak megfelelő adatbázis-táblaszerkezettel lehet megvalósítani.

A `TEMPENABLED` absztrakt tábla az objektumok statikus részét tárolja, míg a `TEMPORAL` tábla a dinamikus részüket. A dinamikus táblából egy külső kulccsal (`FK STAT`) mutatunk a statikus táblába, valamint egy-egy külső kulcs valósítja meg a kettős láncolást is, de ezek a saját táblában található megelőző, ill. következő rekordra mutatnak.

A táblaszerkezet most is absztrakt, vagyis konkrét adatmodell esetében a hasonló működéseket az absztrakt tábla rögzített mezőinek öröklötetésével (beillesztésével) lehet biztosítani.

3.3. Időfüggő modellszint megvalósítása



8. ábra. Időfüggő, dinamikus modellszint táblái

A kétszintű rendszerben időfüggő modellszint megvalósításához a kétféle működést kombinálni kell. Ez azt jelenti, hogy az alkalmazói osztályszerkezetnek örökölnie kell mind a dinamikusan változtatható modellelemek esetében, mind az időfüggést megvalósító modellelemek esetében meghatározott absztrakt osztályoktól.

Az adatbázis szintjén lényegében ugyanez történik: az alábbi ábra bemutatja, hogy hogyan származtathatók a kétszintű időfüggést megvalósító táblák (DYNTEMPGEN, DYNTEMPFEACLASS) a kétszintű dinamikus működést, valamint az időfüggő működést megvalósítókból. Az ábrán a két leszármaztatott táblát álló betűkkel, konkrét táblaként jeleztük. Ez azt fejezi ki, hogy egy időfüggő, dinamikus modellszintet használni kívánó alkalmazás esetében a táblák közvetlen relációs adatbázisbeli megvalósítása javasolt, ahol az egyes mezők az ős-absztrakt táblák definíciójából öröklődéssel állnak elő.

3.4. Dinamikus taxonszerkezet megvalósítása

Természetvédelmi célú rendszerben szükséges a Linné-féle rendszertant leíró ún. „biológiai taxonszerkezet" ábrázolása is. Mivel a tárgyra vonatkozó ismeretek forrongók, ezeket mindenképpen dinamikusan változtatható módon kell megvalósítani. Másrészt viszont, korábbi időpontokra vonatkozó adatok kiértékeléséhez az egésznek időfüggőnek kell lennie, hogy bármilyen korábbi időpont szimulálható legyen. Ez azt jelenti, hogy a taxonokat leíró osztályszerkezet futási időben változhat, és a változást az időben követni kell. Az időfüggő, dinamikus modellszint az alábbi igényekre kézenfekvő megoldást ad.

Megjegyezzük azonban, hogy a fenti, általános elvekből levezetett szerkezet mellett szükséges egy mindezek felett álló „jogfolytonossági" viszony tárolása is. Ez azt írja le, hogy egy taxon jogelődje, ill. jogutódja melyik mások taxon lenne. Az alább részletezett műveletek a szimmetrikus összeolvadást és szétválást is megengedik, tehát a jogfolytonossági viszony $n \circ m$-es többszörösségű. Emiatt egyszerű külső kulccsal nem tárolható valamelyik már meglevő táblában, hanem számára külön tábla felvétele szükséges.

A dinamikus taxonszerkezettel összefüggésben a következő műveletek megvalósítása szükséges:

1. *Taxon átnevezése:* Az eredmény az eredetivel – a megváltozott névtől eltekintve – teljesen megegyezik. A művelet végrehajtása a név módosítását és új időváltozat létrehozását vonja maga után. A jogfolytonossági viszonyban semmilyen változás nem történik.
2. *Taxon átnevezése és más tulajdonságok módosítása:* Az eredmény részben módosult. A művelet végrehajtása során a szóban forgó tulajdonságok módosulnak, és új időváltozat jön létre. A jogfolytonossági viszonyban semmilyen változás nem történik.
3. *Szimmetrikus összeolvadás:* Két taxon összevonása eggyé, új néven. A művelet feltétele, hogy a két taxon szigorúan testvérpár legyen, vagyis közös

közvetlen őstől származzanak. Egyes tulajdonságok megegyezését illetően további feltételek is állíthatók. A két eredeti taxon megszűnik, és tovább nem használhatók, majd létrejön egy új taxon. Az új taxon egyes tulajdonságait kívülről kapja, más tulajdonságai közösek. Ezen túli tulajdonságaira egyesítési megoldást kell meghatározni. A nyitott feltételeket illető meghatározások az elemzési, ill. a tervezési modell részei. Az új taxon jogutódja a két összevont taxonnak.

4. *Aszimmetrikus összeolvadás:* Két taxon összevonása eggyé valamelyik nevén. Taxon beolvadása másik taxonba. A tulajdonságokat illetően az összeolvadásra további feltételek is kiköthetők (ld. feljebb). A megszűnő taxon jogelődje a másiknak.
5. *Szimmetrikus szétválás:* Egy taxon szétszedése két vagy több új taxonra. Több taxonpéldány létrehozása, ahol mindegyik új nevet (és új azonosítót) kap. Az eredeti taxon jogelődje mindegyik új taxonnak.
6. *Aszimmetrikus szétválás:* Egy taxon szétszedése két vagy több taxonra, miközben egy megtartja a korábbi nevét (és azonosítóját). A korábbihoz hasonló művelet. A megmaradó taxon jogelődje többinek.
7. *Szintemelkedés:* Egy alfaj faji szintre emelkedik. Az eredeti taxon megmarad, csupán az alfajt jelző mező lesz fajt jelző értékké átírva. Az alfajhoz vezető általánosítás megszűnik, helyette a saját fajához vezető általánosítással megegyező általánosítás jön létre.

4. Eredmények összegzése és köszönetnyilvánítás

A cikkben leírtak az ún. kétszintű vagy modellvezérelt szoftverek működésének és készítésének néhány alapkérdését, valamint a kidolgozott megoldások alkalmazását mutatták be természetvédelmi célú szoftveralkalmazásokban. A megoldások általánosak, vagyis nem csak a természetvédelemben alkalmazhatók. Bár előtanulmányokban egyes megoldások tesztelve lettek, a leírtak egészének konkrét működés közbeni vizsgálata még nem történt meg. A cikk megírását megelőzte egy természetvédelmi célú információs rendszer tervének rögzítése, de a konkrét megvalósítási tervek elkészítése és az implementációs munkák megkezdése csak a későbbiekben várható.

Köszönetemet szeretném kifejezni a SILK EU projektben résztvevő munkatársaimnak, mert a leírtak kikristályosításában a SILK felépítmény megismerése, és a projektben történő aktív részvételem meghatározó szerepet játszott. Úgyszintén köszönetemet szeretném kifejezni a Természet és Környezetvédelmi Minisztérium Természetvédelmi Főosztályán működtetett Természetvédelmi Informatikai Tanácsadó Testület tagjainak, akik révén a szakmai munka egyáltalán megtörténhetett, ill. a szükséges szakmai információk birtokába juthattam.

Hivatkozások

[1] JOAQUIN MILLER-JISHNU MUKERJI: *Model Driven Architecture*, OMG Document, July 2001.

[2] JON SIEGEL: *Developing in OMG's Model-Driven Architecture*, OMG White Paper, November 2001.

[3] JOAQUIN MILLER-JISHNU MUKERJI: *MDA Guide Version 1.0*, OMG Document, July 2003.

[4] R.G.G.CATTELL-D.BARRY ÉS MÁSOK: *The Object Data Standard: ODMG 3.0*, Morgan Kaufmann publishers San Francisco, USA 1999.

[5] *Human-Usable Textual Notation (HUTN) Specification Version 1.0*, OMG, August 2004.

[6] *SILAN – the SILK language*, IQSOFT, Hungary 2000

[7] *Rumbaugh-Jacobson-Booch*: *Unified Modelling Language Reference Manual*, Addison-Wesley-Longman Inc. 1999.

KILIÁN IMRE
PTE-TTK, Informatika tanszék
7624 Pécs, Ifjúság u. 6.
kilian@gamma.ttk.pte.hu

THE CONSTRUCTION OF MODEL DRIVEN SOFTWARE II.

IMRE KILIÁN

The article demonstrates how and when the application of model driven architecture can be advantageous. In the first section the implementation is described: starting from the simplest, partially model-driven architecture, when only a dynamic property-list is aimed, up to the complete model-level.
The second section describes how the concept of two level software can be used for the dynamic behaviour of taxon hierarchies in biological and ecological software applications. This section also presents the concept and implementation of a 'time machine', i.e. a framework that enables the retrospective operation of a software.

Alkalmazott Matematikai Lapok **26** (2009), 289-315.

A SHAPLEY-ÉRTÉK AXIOMATIZÁLÁSAI[1]

PINTÉR MIKLÓS

A Shapley-érték a koalíciós formában adott játékok egyik legismertebb megoldáskoncepciója. Szokás az irodalomban a Shapley-értéket axiomatikusan jellemezni, leírni. Ebben a cikkben a Shapley-érték négy axiomatizációjával foglalkozunk: a Hart és Mas-Colell-féle potenciállal, Shapley eredeti, a van den Brink-féle és a Young-féle karakterizációkkal. A fenti négy axiomatizálás érvényességét vizsgáljuk az átruházható hasznosságú koalíciós formában adott játékok tizenhat játékosztályán. Eredményeinket egy táblázatban foglaljuk össze.

1. Bevezető

A Shapley-érték a koalíciós formában adott átruházható hasznosságú játékokon[2] (a továbbiakban röviden csak játékok) értelmezett megoldási koncepciók egyik legnépszerűbbike. Mind elméletileg, mind az alkalmazások tekintetében széleskörben használt (az alkalmazásokról Moretti és Patrone [11] cikke ad szisztematikus áttekintést, míg konkrét alkalmazásokra magyar nyelven lásd pl. Csóka [4] és Pintér [17] cikkeket).

Ugyanakkor, az alkalmazó szempontjából is fontos annak megértése, hogy valójában mit is jelent a Shapley-érték. Ez a fajta megértés, jellemzés a tárgya a Shapley-érték különböző axiomatizálásainak. Egy axiomatizálás nem más, mint annak megmutatása, hogy valamely rögzített játékosztályon a Shapley-érték ekvivalens bizonyos axiómákkal. Magyarán szólva, az adott játékosztályon a Shapley-érték egyértelműen jellemezhető az adott tulajdonságokkal (axiómák).

Ahogy fent jeleztük, az egyes axiomatizálások csak adott, rögzített játékosztályok mellett igazak. A kérdés az, hogy milyen játékosztályok mellett mely axiomatizálások érvényesek és melyek nem.

Egyes alkalmazások jól ismert, népszerű játékosztályokhoz köthetőek. Csak példa jelleggel: a nemnegatív szubadditív játékok és a költségjátékok osztályai egybeesnek, szintén megegyezik a nemnegatív nulla-normalizált szuperadditív

[1] Köszönet az anonim bírálónak a megjegyzésekért és az észrevételekért. Ez a cikk az Országos Kutatási és Tudományos Alap (OTKA) pályázata és a Magyar Tudományos Akadémia Bolyai János ösztöndíja támogatásával készült.

[2] A magyar nyelvű irodalomban az átváltható hasznosságú játékok elnevezés is elterjedt.

játékok és a megtakarítási játékok osztálya (lásd a fogalmakra pl. Driessen [5]), ill. ugyancsak fennáll az egybeesés a monoton és a feszítő-hálózatjátékok (spanning network games) játékosztályokra (lásd Van Den Nouweland et al. [15]).

Ebben a cikkben a Shapley-érték négy axiomatizálását vizsgáljuk: (1) a Hart és Mas-Colell-féle [9] potenciált, (2) Shapley eredeti [20] axiomatizálását, melyet később Dubey [6], ill. Peleg és Sudhölter [16] tovább finomított, (3) van den Brink [1] megközelítését és (4) Young [22] karakterizációját. A fenti karakterizációkat tizenhat játékosztályon vizsgáljuk (lásd 2.3. definíció).

A Shapley-érték számos egyéb axiomatizálása ismert az irodalomban, többek között Chun [2], [3], Hart és Mas-Colell [9] redukált játékra épülő karakterizációja, Lange és Grabisch [10], Roth [18]. Ebben a cikkben ahelyett, hogy bevezetünk egy új karakterizációt, négy jól ismert karakterizációt vetünk egybe. Véleményünk szerint a választott karakterizációk, a terjedelmi korlátok figyelembe vétele mellett, jól reprezentálják a Shapley-érték különféle axiomatizációit.

Eredményeinket az 1. táblázat tartalmazza. Három elméleti eredményt szeretnénk hangsúlyozni. (1) Az alapjáték fogalma (lásd a 2.6. definíciót) segítségével Shapley eredeti axiomatizálásának érvényességét olyan esetekben is tudjuk vizsgálni (pl. lényeges játékok, lásd a 4.3. következményt), amikor a korábbi fogalmakkal az nem volt lehetséges. Az alapjátékok fogalma előkerül van den Brink megközelítésének tárgyalásakor is. (2) Olyan módon általánosítjuk van den Brink axiomatizációs tételét (lásd 5.1. tétel), hogy lehetővé válik az adott jellemzés érvényességének vizsgálata minden górcső alá vont játékosztályon (lásd az 5.3., 5.4., 5.5. következményeket). (3) Teljesen új bizonyítást adunk a Young-féle axiomatizálásra (lásd a 6.1. tételt), és ezzel az új bizonyítással megmutatjuk, hogy a Young-féle karakterizáció minden, ebben a cikkben tárgyalt játékosztályon érvényes (lásd a 6.2. következményt).

A cikk felépítése a következő. A 2. részben bevezetjük a cikkben használt jelöléseket és alapfogalmakat. A 3., 4., 5. és 6. részek rendre Hart és Mas-Colell, Shapley eredeti, van den Brink és Young axiomatizálásait tárgyalják. Az utolsó rész az összefoglalásé. Egy hosszú bizonyítást az Appendixbe tettünk.

2. Jelölések, alapfogalmak

Jelölések: tetszőleges N halmaz esetén $|N|$ az N halmaz számossága (ha $x \in \mathbb{R}$, akkor $|x|$ jelentése: x abszolút értéke), $\mathcal{P}(N)$ jelöli az N halmaz összes részhalmazainak osztályát. $A \subset B$ jelentése: $A \subseteq B$ és $A \neq B$. Lin (A) az A lineáris burka, hasonlóan cone (A) a legszűkebb konvex kúp ami tartalmazza A-t.

2.1. Definíció. Legyen $N \neq \emptyset$, $|N| < \infty$, és $v : \mathcal{P}(N) \to \mathbb{R}$ olyan függvény, hogy $v(\emptyset) = 0$. Ekkor N-t, v-t rendre a játékosok halmazának, ill. átruházható hasznosságú koalíciós formában adott játéknak (ezentúl röviden csak játéknak) nevezzük. Továbbá, $\mathcal{G}^N$ jelöli az N játékoshalmazzal rendelkező játékok osztályát.

Megjegyzés. Könnyen látható, hogy $\mathcal{G}^N$ és $\mathbb{R}^{2^{|N|}-1}$ izomorfak. A továbbiakban egy rögzített izomorfizmus[3] mellett feltesszük, hogy $\mathcal{G}^N$ és $\mathbb{R}^{2^{|N|}-1}$ megegyezik.

2.2. Definíció. Legyen $v \in \mathcal{G}^N$ és $i \in N$ tetszőlegesen rögzített, és $\forall S \subseteq N$-re legyen $v_i'(S) \triangleq v(S \cup \{i\}) - v(S)$. Ekkor v_i'-t az i játékos v játékbeli határhozzájárulási függvényének nevezzük.

Tehát $v_i'(S)$ az i játékos v játékbeli határhozzájárulása az S koalícióhoz.

2.3. Definíció. Egy $v \in \mathcal{G}^N$ játék

- lényeges, ha $v(N) > \sum_{i \in N} v(\{i\})$,
- konvex, ha $\forall S, T \subseteq N$-re $v(S) + v(T) \leq v(S \cup T) + v(S \cap T)$,
- szigorúan konvex, ha $\forall S, T \subseteq N$-re, hogy $S \not\subseteq T$, $T \not\subseteq S$: $v(S) + v(T) < v(S \cup T) + v(S \cap T)$,
- szuperadditív, ha $\forall S, T \subseteq N$-re, hogy $S \cap T = \emptyset$: $v(S) + v(T) \leq v(S \cup T)$,
- szigorúan szuperadditív, ha $\forall S, T \subseteq N$-re, hogy $S \cap T = \emptyset$, $S, T \neq \emptyset$: $v(S) + v(T) < v(S \cup T)$,
- gyengén-szuperadditív, ha $\forall S, T \subseteq N$-re, hogy $S \cap T = \emptyset$, $|S| = 1$: $v(S) + v(T) \leq v(S \cup T)$,
- szigorúan gyengén-szuperadditív, ha $\forall S, T \subseteq N$-re, hogy $S \cap T = \emptyset$, $|S| = 1$, $T \neq \emptyset$: $v(S) + v(T) < v(S \cup T)$,
- monoton, ha $\forall S, T \subseteq N$-re, hogy $S \subseteq T$: $v(S) \leq v(T)$,
- szigorúan monoton, ha $\forall S, T \subseteq N$-re, hogy $S \subset T$: $v(S) < v(T)$,
- additív, ha $\forall S, T \subseteq N$-re, hogy $S \cap T = \emptyset$: $v(S) + v(T) = v(S \cup T)$,
- gyengén-szubadditív, ha $\forall S, T \subseteq N$-re, hogy $S \cap T = \emptyset$, $|S| = 1$: $v(S) + v(T) \geq v(S \cup T)$,
- szigorúan gyengén-szubadditív, ha $\forall S, T \subseteq N$-re, hogy $S \cap T = \emptyset$, $|S| = 1$, $T \neq \emptyset$: $v(S) + v(T) > v(S \cup T)$,
- szubadditív, ha $\forall S, T \subseteq N$-re, hogy $S \cap T = \emptyset$: $v(S) + v(T) \geq v(S \cup T)$,
- szigorúan szubadditív, ha $\forall S, T \subseteq N$-re, hogy $S \cap T = \emptyset$, $S, T \neq \emptyset$: $v(S) + v(T) > v(S \cup T)$,
- konkáv, ha $\forall S, T \subseteq N$-re $v(S) + v(T) \geq v(S \cup T) + v(S \cap T)$,
- szigorúan konkáv, ha $\forall S, T \subseteq N$-re, hogy $S \not\subseteq T$, $T \not\subseteq S$: $v(S) + v(T) > v(S \cup T) + v(S \cap T)$.

[3] A rögzített izomorfizmus a következő: vegyünk egy tetszőleges teljes rendezést N-en, tehát feltehetjük, hogy $N = \{1, \ldots, |N|\}$; és $\forall v \in \mathcal{G}^N$-re legyen $v \triangleq (v(\{1\}), \ldots, v(\{|N|\}), v(\{1,2\}), \ldots, v(\{|N|-1, |N|\}), \ldots, v(N)) \in \mathbb{R}^{2^{|N|}-1}$.

Ebben a cikkben a fent bevezetett játékosztályokra koncentrálunk. A következő, az irodalomban jól ismert eredményt bizonyítás nélkül adjuk közre.

2.1. LEMMA. *A $v \in \mathcal{G}^N$ játék pontosan akkor (szigorúan) konvex / (szigorúan) konkáv, ha $\forall i \in N$-re, $\forall T, Z \subseteq N \setminus \{i\}$-re, hogy $Z \subset T$:*

$$v_i'(Z) \leq v_i'(T) (v_i'(Z) < v_i'(T)) / v_i'(Z) \geq v_i'(T) (v_i'(Z) > v_i'(T)).$$

2.4. Definíció. A $v \in \mathcal{G}^N$ játék duálisa az a $\bar{v} \in \mathcal{G}^N$ játék, hogy $\forall S \subseteq N$-re $\bar{v}(S) = v(N) - v(N \setminus S)$.

A következő segédtételben összefoglaljuk a duális játékok néhány nyilvánvaló tulajdonságát.

2.2. LEMMA. *Tekintsük a következő pontokat:*

1. *Legyen $v \in \mathcal{G}^N$ tetszőlegesen rögzített, ekkor $\bar{\bar{v}} = v$.*
2. *Egy (szigorúan) konvex játék duálisa (szigorúan) konkáv játék, ill. egy (szigorúan) konkáv játék duálisa (szigorúan) konvex játék.*

2.5. Definíció. Legyen $v \in \mathcal{G}^N$ tetszőlegesen rögzített. $i \sim^v j$ $(i, j \in N)$, ha $\forall S \subseteq N$-re, hogy $i, j \notin S$: $v_i'(S) = v_j'(S)$. Továbbá, ha $S \subseteq N$ olyan, hogy $\forall i, j \in S$-re $i \sim^v j$, akkor azt mondjuk, hogy S ekvivalenciahalmaz a v játékban.

Könnyen látható, hogy tetszőleges $v \in \mathcal{G}^N$ játékra, $\sim^v$ ekvivalenciareláció $N \times N$-en.

2.6. Definíció. A $v \in \mathcal{G}^N$ játék alapjáték, ha $(i, j \notin NP(v)) \Rightarrow (i \sim^v j)$, ahol $NP(v) \triangleq \{k \in N \mid v_k' = 0\}$.

Magyarán szólva, a v játék alapjáték, ha nem nulla játékosai ekvivalensek.

2.7. Definíció. Legyen N, $T \subseteq N$, $T \neq \emptyset$ tetszőlegesen rögzített, és $\forall S \subseteq N$-re

$$u_T(S) \triangleq \begin{cases} 1, & \text{ha } T \subseteq S \\ 0 & \text{különben.} \end{cases}$$

Az u_T játékot a T koalíción értelmezett egyetértési játéknak nevezzük.

Világos, hogy minden egyetértési játék alapjáték, de nem minden alapjáték egyetértési játék (pl. tetszőleges T-re, αu_T alapjáték, de nem egyetértési játék $(\alpha \neq 1)$). A következő segédtételben, amit bizonyítás nélkül közlünk, összefoglaljuk az alapjátékok néhány nyilvánvaló tulajdonságát.

2.3. LEMMA. *Tekintsük a következő pontokat:*

1. *Ha $v \in \mathcal{G}^N$ alapjáték, akkor $\forall \alpha \in \mathbb{R}$-re αv szintén alapjáték.*

2. *Tetszőleges* $v \in \mathcal{G}^N$*-re ha* $i \in NP(v)$*, akkor* $i \in NP(\bar{v})$.

3. *Tetszőleges* $v \in \mathcal{G}^N$*-re ha* $i \sim^v j$*, akkor* $i \sim^{\bar{v}} j$.

4. *Alapjáték duálisa alapjáték.*

5. *Legyen* $v \triangleq \sum_{i=1}^{k} \alpha_i v_i$*. Ekkor* $\bar{v} = \sum_{i=1}^{k} \alpha_i \bar{v}_i$.

A következő definícióban a megoldás fogalmát vezetjük be.

2.8. Definíció. A $\psi : A \to \mathbb{R}^N$ függvényt, ahol $A \subseteq \mathcal{G}^N$, az A halmazon értelmezett megoldásnak nevezzük.

A 2.8. definícióból kiderül, hogy ebben a cikkben a megoldás egy pontértékű függvény. Mivel a Shapley-érték pontértékű megoldás, így érthető a pontértékűség megszorítás.

2.9. Definíció. (Shapley [20]) Legyen $v \in \mathcal{G}^N$ tetszőlegesen rögzített, és $\forall i \in N$-re legyen

$$\phi_i(v) \triangleq \sum_{S \subseteq N \setminus \{i\}} v_i'(S) \frac{|S|!(|N \setminus S| - 1)!}{|N|!}.$$

Ekkor $\phi_i(v)$-t az i játékos v játékbeli Shapley-értékének nevezzük. A továbbiakban jelölje ϕ a Shapley-megoldást.

A következőkben bevezetjük a cikkben tárgyalt axiómákat.

2.10. Definíció. ψ az $A \subseteq \mathcal{G}^N$ halmazon értelmezett megoldás

- Pareto-optimális (Pareto optimal / PO), ha $\forall v \in A$-ra $\sum_{i \in N} \psi_i(v) = v(N)$,
- nulla játékos tulajdonságú (null player property / NP), ha $\forall v \in A$-ra, $\forall i \in N$-re $(v_i' = 0) \Rightarrow (\psi_i(v) = 0)$,
- egyenlően kezelő (equal treatment property / ETP), ha $\forall v \in A$-ra $(i \sim^v j) \Rightarrow (\psi_i(v) = \psi_j(v))$,
- additív (additive / ADD), ha $\forall v, w \in A$-ra, hogy $v + w \in A$: $\psi(v + w) = \psi(v) + \psi(w)$,
- fair tulajdonságú (fairness property / FP), ha $\forall v, w \in A$-ra $\forall i, j \in N$-re, hogy $v + w \in A$ és $i \sim^w j$: $\psi_i(v + w) - \psi_i(v) = \psi_j(v + w) - \psi_j(v)$,
- egyenlőség monoton (equal marginality property / EMP), ha $\forall v, w \in A$-ra $(v_i' = w_i') \Rightarrow (\psi_i(v) = \psi_i(w))$.

A következő segédtétel az FP, ill. ETP és ADD tulajdonságok közötti kapcsolatot jellemzi.

2.4. LEMMA. *Ha a* ψ *megoldás* ETP *és* ADD*, akkor* FP *is.*

Bizonyítás. Lásd van den Brink [1] Proposition 2.3. (i) pont 311. old. □

A következő segédtétel, aminek bizonyítását az olvasóra bízzuk, az alapjáték fogalom erejét és értelmét illusztrálja.

2.5. LEMMA. *Legyen $v \in \mathcal{G}^N$ tetszőlegesen rögzített alapjáték. A ψ megoldás pontosan akkor PO, NP és ETP, ha $\psi(v) = \phi(v)$.*

A következő eredmény jól ismert az irodalomban, így eltekintünk bizonyításától.

2.1. ÁLLÍTÁS. *A Shapley-megoldás PO, NP, ETP, ADD, FP és EMP.*

3. A potenciál (Hart és Mas-Colell)

Ebben a részben Hart és Mas-Colell [9] potenciálfüggvényen alapuló Shapley-érték karakterizációját tárgyaljuk.

3.1. Definíció. Legyen $v \in \mathcal{G}^N$ és $T \subseteq N$, $T \neq \emptyset$ tetszőlegesen rögzített. Ekkor a v játék T-n értelmezett részjátéka $v^T \in \mathcal{G}^T$ a következő: $\forall S \subseteq T$-re

$$v^T(S) = v(S).$$

Világos, hogy v^T-t csak T részhalmazain kell definiálni.

3.2. Definíció. Legyen $A \subseteq \Gamma^N \triangleq \bigcup_{T \subseteq N,\ T \neq \emptyset} \mathcal{G}^T$, $P : A \to \mathbb{R}$, és $\forall v \in \mathcal{G}^T \cap A$-ra, $\forall i \in T$-re, hogy $|T| = 1$ vagy $v^{T\setminus\{i\}} \in A$:

$$P_i'(v) \triangleq \begin{cases} P(v), & \text{ha } |T| = 1 \\ P(v) - P(v^{T\setminus\{i\}}) & \text{különben.} \end{cases} \tag{1}$$

Továbbá, ha $\forall v \in \mathcal{G}^T \cap A$-ra, hogy $|T| = 1$ vagy $\forall i \in T$-re $v^{T\setminus\{i\}} \in A$:

$$\sum_{i \in T} P_i'(v) = v(T),$$

akkor P-t az A halmazon értelmezett potenciálnak nevezzük.

3.3. Definíció. Az $A \subseteq \Gamma^N$ halmaz részjáték zárt, ha $\forall T \subseteq N$-re, hogy $|T| > 1$, $\forall v \in \mathcal{G}^T \cap A$-ra, $\forall i \in T$-re $v^{T\setminus\{i\}} \in A$.

Mivel nincsen játék játékos nélkül, azaz a játékoshalmaz nemüres, ezért a részjáték fogalmára csak akkor támaszkodunk, ha legalább két játékos van T-ben.

3.1. TÉTEL. *Legyen $A \subseteq \Gamma^N$ egy részjáték zárt játékosztály. Ekkor P az A halmazon értelmezett függvény pontosan akkor potenciál, ha $\forall v \in \mathcal{G}^T \cap A$-ra és $\forall i \in T$-re $P_i'(v) = \phi_i(v)$.*

Bizonyítás. Lásd Peleg és Sudhölter [16] Theorem 8.4.4. (216-217 old.). □

A következőkben a korábban bevezetett játékosztályokat vesszük górcső alá.

3.1. KÖVETKEZMÉNY. *P a (szigorúan) konvex / (szigorúan) szuperadditív / (szigorúan) gyengén-szuperadditív / (szigorúan) monoton / additív / (szigorúan) gyengén-szubadditív / (szigorúan) szubadditív / (szigorúan) konkáv játékok osztályán értelmezett függvény pontosan akkor potenciál, ha $\forall v \in \mathcal{G}^T$ (szigorúan) konvex / (szigorúan) szuperadditív / (szigorúan) gyengén-szuperadditív / (szigorúan) monoton / additív / (szigorúan) gyengén-szubadditív / (szigorúan) szubadditív / (szigorúan) konkáv játékra és $\forall i \in T$-re $P_i'(v) = \phi_i(v)$.*

Bizonyítás. Könnyen látható, hogy a (szigorúan) konvex / (szigorúan) szuperadditív / (szigorúan) gyengén-szuperadditív / (szigorúan) monoton / additív / (szigorúan) gyengén-szubadditív / (szigorúan) szubadditív / (szigorúan) konkáv játékok osztálya részjáték zárt, tehát alkalmazhatjuk a 3.1. tételt. □

3.2. KÖVETKEZMÉNY. *A lényeges játékok osztályán van olyan P potenciál, hogy $\exists v \in \mathcal{G}^T$ olyan lényeges játék, hogy $\exists i \in T$: $P_i'(v) \neq \phi_i(v)$.*

Bizonyítás. Legyen $N \triangleq \{1,2\}$, $v \in \mathcal{G}^N$ egy tetszőleges lényeges játék. Ekkor sem $v^{N\setminus\{1\}}$, sem $v^{N\setminus\{2\}}$ nem lényeges játék. Általában, egy a lényeges játékok osztályán értelmezett potenciál nem jól definiált a két játékossal rendelkező lényeges játékokon. Mivel a potenciál rekurzióval definiált (lásd a 3.2. definíciót), így annak értéke a két játékossal rendelkező lényeges játékon vett értékektől függ. Tehát kontinuum sok potenciál van a lényeges játékok osztályán. □

4. A Shapley-féle jellemzés

Shapley [20] eredeti axiomatizációjával foglalkozunk ebben a részben. A következő tétel az egyre inkább letisztázott, finomított tételek és bizonyítások – Shapley, Dubey [6], Peleg és Sudhölter [16] – sorába illeszkedik.

4.1. TÉTEL. *Legyen $A \subseteq \mathcal{G}^N$ olyan, hogy $\forall v \in A$-hoz $\exists v_1, \ldots, v_k \in A$ alapjáték, hogy*

1. $\operatorname{cone}(\{v_i\}_{i=1}^k) \setminus \{0\} \subseteq A$,
2. $v \in \operatorname{Lin}(\{v_i\}_{i=1}^k)$.

Ekkor az A-n értelmezett megoldás ψ pontosan akkor PO, NP, ETP és ADD, ha $\psi = \phi$.

Bizonyítás.
Szükséges: Lásd a 2.1. állítást.

Elégséges: Legyen $v \in A$ egy tetszőlegesen rögzített játék, és ψ az A-n értelmezett PO, NP, ETP és ADD megoldás. Ha $v = 0$, akkor a PO és ETP tulajdonságok miatt $\psi(v) = \phi(v)$.

Tegyük fel, hogy $v \neq 0$. A 2. pontból $\exists \alpha_1, \ldots, \alpha_k \in \mathbb{R} \setminus \{0\}$, hogy

$$v = \sum_{i=1}^{k} \alpha_i v_i.$$

Legyen $Neg \triangleq \{i \in \{1, \ldots, k\} \mid \alpha_i < 0\}$. Az 1. pont miatt

$$\left(- \sum_{i \in Neg} \alpha_i v_i \right) \in A,$$

és

$$\left(\sum_{i \in \{1,\ldots,k\} \setminus Neg} \alpha_i v_i \right) \in A.$$

Továbbá

$$v + \left(- \sum_{i \in Neg} \alpha_i v_i \right) = \sum_{i \in \{1,\ldots,k\} \setminus Neg} \alpha_i v_i.$$

A 2.3., 2.5. segédtételek és ADD miatt

$$\psi \left(- \sum_{i \in Neg} \alpha_i v_i \right) = \phi \left(- \sum_{i \in Neg} \alpha_i v_i \right),$$

és

$$\psi \left(\sum_{i \in \{1,\ldots,k\} \setminus Neg} \alpha_i v_i \right) = \phi \left(\sum_{i \in \{1,\ldots,k\} \setminus Neg} \alpha_i v_i \right),$$

így a 2.1. állítás és az ADD tulajdonság miatt

$$\psi(v) = \phi(v).$$

□

A 4.1. tétel Peleg és Sudhölter tételének egy általánosítása. A két tétel közötti különbség „csak" annyi, hogy míg Peleg és Sudhölter az egyetértési játékok által kifeszített konvex kúppal dolgozik, addig mi tetszőleges alapjátékok által kifeszített kúpot használunk.

4.1. KÖVETKEZMÉNY. *ψ a konvex / szuperadditív / gyengén-szuperadditív / monoton / additív / gyengén-szubadditív / szubadditív / konkáv játékok osztályán értelmezett megoldás pontosan akkor PO, NP, ETP és ADD, ha $\psi = \phi$.*

Bizonyítás. A konvex / szuperadditív / gyengén-szuperadditív / monoton játékok osztálya tartalmazza cone $(\{u_T\}_{T\subseteq N,\ T\neq\emptyset})$-t az egyetértési játékok által kifeszített kúpot. $\{u_T\}_{T\subseteq N,\ T\neq\emptyset}$ bázisa $\mathbb{R}^{2^{|N|}-1}$-nek (lásd pl. Peleg és Sudhölter Lemma 8.1.4. 203–204 old.), így alkalmazhatjuk a 4.1. tételt.

Az additív játékok osztálya egybeesik Lin $(\{u_T\}_{T\subseteq N,\ |T|=1})$-vel, tehát a 4.1. tétel ebben az esetben is alkalmazható.

A gyengén-szubadditív / szubadditív / konkáv játékok osztálya tartalmazza cone $(\{\bar{u}_T\}_{T\subseteq N,\ T\neq\emptyset})$-t, az egyetértési játékok duálisai által kifeszített konvex kúpot. A 2.3. segédtétel miatt $\{\bar{u}_T\}_{T\subseteq N,\ T\neq\emptyset}$ bázisa $\mathbb{R}^{2^{|N|}-1}$-nek, tehát a 4.1. tételt alkalmazhatjuk ebben az esetben is. □

Megjegyzés. Vegyük észre, hogy az additív játékok osztályán a PO és NP tulajdonságokból következik ETP, tehát a 4.1. következményt újrafogalmazhatjuk a következő formában:

„ψ az additív játékok osztályán értelmezett megoldás pontosan akkor PO, NP és ADD, ha $\psi = \phi$."

4.2. KÖVETKEZMÉNY. *A legalább két játékossal rendelkező ($|N| > 1$) szigorúan konvex / szigorúan szuperadditív / szigorúan gyengén-szuperadditív / szigorúan monoton / szigorúan gyengén-szubadditív / szigorúan szubadditív / szigorúan konkáv játékok osztályán van olyan PO, NP, ETP és ADD megoldás ψ, hogy $\psi \neq \phi$.*

Bizonyítás. Legyen $\forall v \in \mathcal{G}^N$ tetszőleges szigorúan konvex / szigorúan szuperadditív / szigorúan gyengén-szuperadditív / szigorúan monoton / szigorúan gyengén-szubadditív / szigorúan szubadditív / szigorúan konkáv játék, és $\forall i \in N$-re legyen $\psi_i(v) \triangleq \dfrac{v(N)}{|N|}$ (egalitáriánus megoldás). Világos, hogy $\psi \neq \phi$.

Könnyen látható, hogy ψ rendelkezik az NP (nincsen nulla játékos ezekben „szigorú" játékosztályokban), PO, ETP és ADD tulajdonságokkal. □

Megjegyzés. Világos, hogy ha $|N| = 1$, akkor tetszőleges játék estén a PO tulajdonság egyedül biztosítja, hogy $\psi = \phi$.

Vegyük észre, hogy a 4.1. következmény nem támaszkodik a 4.1. tétel teljes „erejére". Tulajdonképpen Peleg és Sudhölter eredményének egy „duál verziója" is elég a 4.1. következmény bizonyításához. A következőkben egy olyan eredményt mutatunk be, ami már nem látható be Peleg és Sudhölter tételével, tehát a következő eredmény azt mutatja, hogy az általánosításunk releváns, és új eredményt hoz.

4.3. KÖVETKEZMÉNY. *Ha $|N| = 2$, akkor van olyan a lényeges játékok osztályán értelmezett PO, NP, ETP és ADD megoldás ψ, hogy $\psi \neq \phi$. Ha azonban $|N| \neq 2$, akkor a PO, NP, ETP és ADD axiómák a lényeges játékok osztályán jellemzik a Shapley-értéket.*

Bizonyítás.

$|N| = 2$: Ebben ez esetben a szigorúan szuperadditív játékok osztálya és a lényeges játékok osztálya egybeesik. Tehát a 4.2. következményből következik az állítás.

$|N| \neq 2$: Feltehetjük, hogy $|N| > 2$. $\forall i \in N$-re legyen

$$v_i(S) \triangleq \begin{cases} 0, & \text{ha } S = \emptyset \text{ vagy } S = \{i\} \\ 1, & \text{ha } |S \setminus \{i\}| = 1 \\ |N| & \text{különben.} \end{cases}$$

Ekkor v_i-k lényeges alapjátékok ($NP(v_i) = \{i\}$), és tetszőleges $|T| > 1$-re u_T szintén lényeges alapjáték. Továbbá, cone $(\{v_i\}_{i\in N} \cup \{u_T\}_{|T|>1}) \setminus \{0\}$ benne van a lényeges játékok osztályában, és $\{v_i\}_{i\in N} \cup \{u_T\}_{|T|>1}$ bázisa $\mathbb{R}^{2^{|N|}-1}$-nek, így alkalmazhatjuk a 4.1. tételt. □

5. van den Brink jellemzése

Ebben a részben a Shapley-érték van den Brink-féle [1] axiomatizálásával foglalkozunk.

5.1. Definíció. Legyen $A \subseteq \mathcal{G}^N$ játékosztály és ψ az A-n értelmezett megoldás tetszőlegesen rögzített. Azt mondjuk, hogy A passzol ψ-hez, ha $\forall v \in A$-ra, hogy $i, j \in N$, $i \sim^v j$: $\exists w \in A$, hogy $i \sim^w j$, $v + w \in A$ és $\psi_i(w) = \psi_j(w)$.

A következő segédtétel a fent bevezetett fogalom „indoklásának" tekinthető.

5.1. LEMMA. *Legyen $A \subseteq \mathcal{G}^N$ játékosztály és ψ A-n értelmezett megoldás olyan, hogy A passzol ψ-hez. Ekkor, ha ψ FP, akkor ETP is.*

Bizonyítás. Legyen $v, w \in \mathcal{G}^N$ tetszőlegesen rögzített úgy, ahogy az 5.1. definícióban szerepelnek. Az FP tulajdonságból

$$\psi_i(v + w) - \psi_i(w) = \psi_j(v + w) - \psi_j(w),$$

így $\psi_i(v + w) = \psi_j(v + w)$. FP miatt

$$\psi_i(v + w) - \psi_i(v) = \psi_j(v + w) - \psi_j(v).$$

Ekkor $\psi_i(v + w) = \psi_j(v + w)$-ből következik, hogy

$$\psi_i(v) = \psi_j(v).$$

□

Van den Brink eredménye (Proposition 2.3. (ii) pont 311. old.) közvetlenül következik a fenti segédtételből.

5.1. KÖVETKEZMÉNY. *Legyen $A \subseteq \mathcal{G}^N$ olyan, hogy $0 \in A$, és ψ az A-n értelmezett megoldás NP és FP. Ekkor ψ ETP.*

Bizonyítás. Legyen $w \triangleq 0$, ekkor NP miatt $\psi(0) = 0$, így A passzol ψ-hez, tehát alkalmazhatjuk az 5.1. segédtételt. □

A következőkben az 5.1. definícióban bevezetett fogalom hasznosságát mutatjuk meg.

5.2. Definíció. Legyen $B \subseteq \mathcal{G}^N$ alapjátékok halmaza tetszőlegesen rögzített. Ha $\forall S \subset N$-re, hogy $|S| = 2$: $\exists v \in B$, hogy $S \subseteq NP(v)$, akkor B-t az alapjátékok ETP-típusú halmazának nevezzük.

Vegyük észre, hogy az egyetértési játékok halmaza, ill. az egyetértési játékok duálisai alkotta halmaz és a 4.3. következményben alkalmazott alapjátékok halmaza legalább négy játékos esetén az alapjátékok ETP-típusú halmazai.

5.2. LEMMA. *Legyen $B \subseteq \mathcal{G}^N$ alapjátékok ETP-típusú halmaza, és ψ a* cone $(B) \setminus \{0\}$ *halmazon értelmezett NP megoldás. Ekkor* cone $(B) \setminus \{0\}$ *passzol ψ-hez.*

Bizonyítás. Legyen $v \in$ cone $(B) \setminus \{0\}$ olyan, hogy $i \sim^v j$ tetszőlegesen rögzített, $w \in B$ olyan, hogy $\{i, j\} \subseteq NP(w)$. Ekkor $v + w \in$ cone $(B) \setminus \{0\}$, ψ NP, így $\psi_i(w) = \psi_j(w)$. □

Összefoglalva a következő eredményre jutunk.

5.1. ÁLLÍTÁS. *Legyen $B \subseteq \mathcal{G}^N$ alapjátékok ETP-típusú halmaza, és ψ* cone $(B) \setminus \{0\}$*-n értelmezett NP megoldás. Ekkor, ha ψ FP, akkor ETP is.*

Bizonyítás. Lásd az 5.1. és 5.2. segédtételeket. □

A következőkben egy újabb fontos fogalmat vezetünk be.

5.3. Definíció. Legyen $B \subseteq \mathcal{G}^N$ alapjátékok halmaza tetszőlegesen rögzített. Ha $\forall v \in B$-re, hogy $2 \leq |NP(v)| < |N|$: $\exists i \in NP(v)$ és $\exists j \notin NP(v)$, hogy $v \circ \pi_{ij} \in$ cone $(B) \setminus \{0\}$, ahol $\pi_{ij} : N \to N$ olyan, hogy

$$\pi_{ij}(x) \triangleq \begin{cases} x, & \text{ha } x \notin \{i, j\} \\ i, & \text{ha } x = j \\ j, & \text{ha } x = i \end{cases},$$

akkor B-t az alapjátékok ADD-típusú halmazának nevezzük.

Vegyük észre, hogy az egyetértési játékok halmaza, ill. az egyetértési játékok duálisai alkotta halmaz és a 4.3. következményben alkalmazott alapjátékok halmaza az alapjátékok ADD-típusú halmazai.

A következő segédtétel bizonyítását az olvasóra bízzuk.

5.3. LEMMA. *Legyen $v \in \mathcal{G}^N$ és $i, j \in N$ tetszőlegesen rögzített. Ekkor*

$$i \sim^{v+v\circ\pi_{ij}} j.$$

A következő állítás matematikai értelemben a fő eredménye ennek résznek.

5.2. ÁLLÍTÁS. *Legyen $B \subseteq \mathcal{G}^N$ alapjátékok ADD-típusú halmaza, és ψ cone $(B) \setminus \{0\}$-n értelmezett olyan PO, FP megoldás, amely tetszőlegesen rögzített az értelmezési tartományában lévő alapjátékokon. Ekkor ψ jóldefiniált, azaz egyértelműen meghatározott.*

Bizonyítás. Legyen $v \in$ cone $(B) \setminus \{0\}$ tetszőlegesen rögzített. Ekkor $v = \sum_{u \in B} \alpha_u u$, és legyen $I(v) \triangleq \{u \in B \mid \alpha_u > 0\}$. $|I(v)|$-n való teljes indukcióval bizonyítunk.

$|I(v)| = 1$: Ekkor $\exists u \in B$, hogy $v = \alpha_u u$, így v alapjáték és $\psi(v)$ jóldefiniált.

$|I(v)| > 1$: Tegyük fel, hogy valamely $1 \leq k < |I(v)|$-re, $\forall A \subseteq I(v)$-re, hogy $|A| \leq k$: $\psi(\sum_{u \in A} \alpha_u u)$ jól definiált. Legyen $C \subseteq I(v)$ olyan tetszőlegesen rögzített halmaz, hogy $|C| = k + 1$, és legyen $z \triangleq \sum_{u \in C} \alpha_u u$.

1. eset: $\exists u_1, u_2 \in C$, hogy $\exists i^*, j^* \in N$: $i^* \sim^{u_1} j^*$, de $i^* \nsim^{u_2} j^*$. Ekkor FP, és $z - \alpha_{u_1} u_1, z - \alpha_{u_2} u_2 \in$ cone $(B) \setminus \{0\}$ következtében $\forall i \in N \setminus \{i^*\}$-ra, hogy $i \sim^{\alpha_{u_2} u_2} i^*$:

$$\psi_{i^*}(z) - \psi_{i^*}(z - \alpha_{u_2} u_2) = \psi_i(z) - \psi_i(z - \alpha_{u_2} u_2), \tag{2}$$

és $\forall j \in N \setminus \{j^*\}$-ra, hogy $j \sim^{\alpha_{u_2} u_2} j^*$:

$$\psi_{j^*}(z) - \psi_{j^*}(z - \alpha_{u_2} u_2) = \psi_j(z) - \psi_j(z - \alpha_{u_2} u_2), \tag{3}$$

és

$$\psi_{i^*}(z) - \psi_{i^*}(z - \alpha_{u_1} u_1) = \psi_{j^*}(z) - \psi_{j^*}(z - \alpha_{u_1} u_1). \tag{4}$$

Továbbá, PO miatt

$$\sum_{i \in N} \psi_i(z) = z(N). \tag{5}$$

Az indukciós hipotézis miatt a (2), (3), (4), (5) lineáris egyenletrendszerben $|N|$ ismeretlen ($\psi_i(z)$, $i \in N$) és $|N|$ egyenlet van, és az egyenletrendszernek egyetlen megoldása van. Tehát $\psi(z)$ jóldefiniált.

2. eset: $\forall u_1, u_2 \in C$-re $NP(u_1) = NP(u_2)$ vagy $NP(u_1) = \complement NP(u_2)$. Ha $\forall u_1, u_2 \in C$-re $NP(u_1) = NP(u_2)$, akkor z alapjáték, így $\psi(z)$ jóldefiniált.

$\exists u_1, u_2 \in C$, hogy $NP(u_1) = \complement NP(u_2)$. Ha $|N| = 2$, akkor feltehetjük, hogy $C = \{u_1, u_2\}$, így $\exists \beta \in \mathbb{R}$, hogy $u_2 = \beta u_1 \circ \pi_{ij}$, ahol $N = \{i, j\}$. Tehát

$z = \alpha_1 u_1 + \alpha_2\beta u_1 \circ \pi_{ij} = m(u_1 + u_1 \circ \pi_{ij}) + (\alpha_1 - m)u_1 + (\alpha_2\beta - m)u_1 \circ \pi_{ij}$, ahol $m = \min\{\alpha_1, \alpha_2\beta\}$. Ekkor az 5.3. segédtétel és az 1. eset miatt $\psi(z)$ jóldefiniált.

Ha $|N| \neq 2$, akkor feltehetjük, hogy $|NP(u_1)| \geq 2$. C két diszjunkt halmazra bontható:

$$U_1 \triangleq \{u \in C \mid NP(u) = NP(u_1)\} \quad \text{és} \quad U_2 \triangleq C \setminus U_1.$$

cone $(B) \setminus \{0\}$ tartalmazza $\sum_{u \in U_1} \alpha_u u$-t és $\sum_{u \in U_2} \alpha_u u$-t, tehát az indukciós hipotézis miatt

$$\psi\left(\sum_{u \in U_1} \alpha_u u\right) \quad \text{és} \quad \psi\left(\sum_{u \in U_2} \alpha_u u\right)$$

jóldefiniált.

B alapjátékok ADD-típusú halmaza, így $\exists i^* \in NP(u_1)$ és $\exists j^* \notin NP(u_1)$, hogy $u_1 \circ \pi_{i^*j^*} \in$ cone $(B) \setminus \{0\}$. Ekkor z, $\alpha_{u_1}u_1$, $\alpha_{u_2}u_2$, i^*, j^* helyére rendre $z + \alpha_{u_1}u_1 \circ \pi_{i^*j^*}$-t, $\alpha_{u_1}(u_1 + u_1 \circ \pi_{i^*j^*})$-t, $\sum_{u \in U_2} \alpha_u u$-t, i^*-t, j^*-t írhatunk a (2), (3), (4), (5) egyenlőségekben. Így az 5.3. segédtétel, az indukciós hipotézis és az 1. eset miatt, ha $|U_2| = 1$, akkor

$$|I(z + \alpha_{u_1}u_1 \circ \pi_{i^*j^*} - \sum_{u \in U_2} \alpha_u u)| = k + 1,$$

de az 1. esetből

$$\psi\left(z + \alpha_{u_1}u_1 \circ \pi_{i^*j^*} - \sum_{u \in U_2} \alpha_u u\right)$$

jóldefiniált: $\psi(z + \alpha_{u_1}u_1 \circ \pi_{i^*j^*})$ jóldefiniált.

Ekkor z, $z - \alpha_{u_1}u_1$, $\alpha_{u_2}u_2$, i^*, j^* helyére rendre z-t, $z + \alpha_{u_1}u_1 \circ \pi_{i^*j^*}$-t, $\sum_{u \in U_2} \alpha_u u$-t, i'-t, hogy $i' \sim^{u_1} i^*$ és $i' \neq i^*$ tetszőlegesen rögzített ($|NP(u_1)| \geq 2$), j^*-t ($i' \sim^{\alpha_{u_1}u_1 \circ \pi_{i^*j^*}} j^*$) írhatunk a (2), (3), (4), (5) egyenlőségekben, és azt kapjuk, hogy $\psi(z)$ jóldefiniált.

Tehát $\psi(v)$ jóldefiniált. □

5.3. Állítás. *Legyen $B \subseteq \mathcal{G}^N$ alapjátékok ADD-típusú halmaza, és ψ* cone $(B) \setminus \{0\}$*-n értelmezett PO, NP és ETP megoldás. Ekkor ψ pontosan akkor FP, ha ADD.*

Bizonyítás.

Szükséges: Lásd a 2.4. segédtételt.

Elégséges: A 2.5. segédtétel miatt ψ jóldefiniált a cone $(B) \setminus \{0\}$ halmazbeli alapjátékokon. Az 5.2. állítás következtében ψ jóldefiniált cone $(B) \setminus \{0\}$-n. Ekkor a 2.4. segédtételből, ha ψ ETP és ADD, akkor FP is, így a jóldefiniált ψ ADD. □

A következő tétel – ami van den Brink fő eredményének (Theorem 2.5. 311–315. old.) általánosítása – ennek a résznek a fő eredménye.

5.1. TÉTEL. *Legyen $A \subseteq \mathcal{G}^N$ olyan, hogy $\forall v \in A$-ra $\exists B \subseteq A$, hogy*

1. $\text{cone}\,(B) \setminus \{0\} \subseteq A$,
2. *B alapjátékok ETP-típusú halmaza,*
3. *B alapjátékok ADD-típusú halmaza,*
4. *$\exists w \in A$ olyan alapjáték, hogy $NP(w) \subseteq NP(v)$, és $v+w \in \text{cone}\,(B)\setminus\{0\}$ vagy $v - w \in \text{cone}\,(B) \setminus \{0\}$.*

Ekkor az A-n értelmezett megoldás ψ pontosan akkor PO, NP és FP, ha $\psi = \phi$.

Bizonyítás.
Szükséges: Lásd a 2.1. állítást.
Elégséges: A $v-w \in \text{cone}\,(B)\setminus\{0\}$ esetet bizonyítjuk, a másik eset bizonyítása teljesen analóg módon megy.
$v - w \in \text{cone}\,(B) \setminus \{0\}$, így a 2.1., 5.1., 5.2. és 5.3. állítások miatt

$$\psi(v - w) = \phi(v - w).$$

Legyen $i^* \in \mathbb{C}NP(v)$ tetszőlegesen rögzített. $NP(w) \subseteq NP(v)$, NP, FP és PO miatt $\forall i \in \mathbb{C}NP(v) \setminus \{i^*\}$-ra

$$\psi_{i^*}(v) - \phi_{i^*}(v - w) = \psi_i(v) - \phi_i(v - w), \tag{6}$$

$\forall i \in NP(v)$-re

$$\psi_i(v) = 0, \tag{7}$$

és

$$\sum_{i \in N} \psi_i(v) = v(N). \tag{8}$$

A (6), (7), (8), (5) lineáris egyenletrendszerben $|N|$ ismeretlen ($\psi_i(v)$, $i \in N$) és $|N|$ egyenlet van, és az egyenletrendszernek egyetlen megoldása van. Tehát $\psi(v)$ jóldefiniált, így a 2.1. állítás következtében $\psi(v) = \phi(v)$. □

Vegyük észre, hogy a 4.1. és 5.1. tételek néhány fontos esetben ekvivalensek.

5.2. KÖVETKEZMÉNY. *Legyen $B \subseteq \mathcal{G}^N$ olyan, hogy*

1. *B alapjátékok ETP-típusú halmaza,*
2. *B alapjátékok ADD-típusú halmaza.*

Továbbá, legyen ψ $\text{cone}\,(B) \setminus \{0\}$-n értelmezett PO és NP megoldás. Ekkor ψ pontosan akkor FP, ha ETP és ADD.

Bizonyítás. Lásd a 2.4. segédtételt és az 5.1., 5.3. állításokat. □

A következőkben az ebben a cikkben tárgyalt játékosztályokat vizsgáljuk. (Természetesen a 4.2. következményt követő megjegyzés erre a karakterizációra is érvényes.)

5.3. KÖVETKEZMÉNY. *ψ a konvex / szuperadditív / gyengén-szuperadditív / monoton / additív / gyengén-szubadditív / szubadditív / konkáv játékok osztályán értelmezett megoldás pontosan akkor PO, NP és FP, ha $\psi = \phi$.*

Bizonyítás. Feltehetjük, hogy $|N| > 1$.

(1) Legyen $v \in \mathcal{G}^N$ tetszőlegesen rögzített konvex / szuperadditív / gyengén-szuperadditív / monoton játék. Legyen B az egyetértési játékok halmaza,

$$v = \sum_{T \subseteq N,\ T \neq \emptyset} \alpha_T u_T, \qquad \alpha \triangleq \max\{-\min_T \alpha_T, 0\}, \qquad w \triangleq (\alpha + 1) \sum_{T \subseteq N,\ T \neq \emptyset} u_T.$$

Világos, hogy w konvex alapjáték és $NP(w) = \emptyset$, $v + w \in \text{cone}\,(B) \setminus \{0\}$, B alapjátékok ETP-típusú és ADD-típusú halmaza, a konvex / szuperadditív / gyengén-szuperadditív / monoton játékok osztálya tartalmazza cone $(B) \setminus \{0\}$-t, így alkalmazhatjuk az 5.1. tételt.

(2) Legyen $v \in \mathcal{G}^N$ tetszőlegesen rögzített additív játék. Legyen

$$B \triangleq \{u_T\}_{T \subseteq N,\ |T|=1}, \qquad v = \sum_{T \subseteq N,\ |T|=1} \alpha_T u_T, \qquad \alpha \triangleq \max\{-\min_T \alpha_T, 0\},$$
$$w \triangleq (\alpha + 1) \sum_{T \subseteq N,\ |T|=1} u_T.$$

Világos, hogy w additív alapjáték és $NP(w) = \emptyset$, $v+w \in \text{cone}\,(B)\setminus\{0\}$, B alapjátékok ETP-típusú és ADD-típusú halmaza, az additív játékok osztálya tartalmazza cone $(B) \setminus \{0\}$-t, így alkalmazhatjuk az 5.1. tételt.

(3) Legyen $v \in \mathcal{G}^N$ tetszőlegesen rögzített gyengén-szubadditív / szubadditív / konkáv játék. Legyen B az egyetértési játékok duálisai alkotta halmaz,

$$v = \sum_{T \subseteq N,\ T \neq \emptyset} \alpha_T \bar{u}_T, \qquad \alpha \triangleq \max\{-\min_T \alpha_T, 0\}, \qquad w \triangleq (\alpha + 1) \sum_{T \subseteq N,\ T \neq \emptyset} \bar{u}_T.$$

Világos, hogy w konkáv alapjáték és $NP(w) = \emptyset$, $v + w \in \text{cone}\,(B) \setminus \{0\}$, B alapjátékok ETP-típusú és ADD-típusú halmaza, a gyengén-szubadditív / szubadditív / konkáv játékok osztálya tartalmazza cone $(B) \setminus \{0\}$-t, így alkalmazhatjuk az 5.1. tételt. □

5.4. KÖVETKEZMÉNY. *Van olyan a legalább két játékossal rendelkező ($|N| \geq 2$) szigorúan konvex / szigorúan szuperadditív / szigorúan gyengén-szuperadditív / szigorúan monoton / szigorúan gyengén-szubadditív / szigorúan szubadditív / szigorúan konkáv játékok osztályán értelmezett PO, NP és FP megoldás ψ, hogy $\psi \neq \phi$.*

Bizonyítás. Lásd a 4.2. következményt. □

5.5. KÖVETKEZMÉNY. *Ha $|N| = 2,3$, akkor van olyan a lényeges játékok osztályán értelmezett PO, NP és FP megoldás ψ, hogy $\psi \neq \phi$. Azonban, ha $|N| \neq 2,3$, akkor a lényeges játékok osztályán a PO, NP, FP axiómák jellemzik a Shapley-értéket.*

Bizonyítás. Ha $|N| \neq 3$, akkor lásd a 4.3. következményt és az 5.3. következmény bizonyítását.

$|N| = 3$: Tekintsük a 4.3. következménybeli alapjátékokat. Legyen

$$N \triangleq \{i_1, i_2, i_3\}, \qquad \psi(v_{i_1}) \triangleq (0,3,0), \qquad \psi(v_{i_2}) \triangleq \psi(v_{i_3}) \triangleq (3,0,0),$$
$$\psi(u_{\{i_1,i_2\}}) \triangleq \psi(u_{\{i_1,i_3\}}) \triangleq \psi(u_N) \triangleq (1,0,0), \quad \psi(u_{\{i_2,i_3\}}) = (0,1,0).$$

Ekkor az 5.2. állítás és az 5.1. tétel bizonyításabeli módszer (5.3. következmény) alkalmazásával ψ egyértelműen kiterjeszthető a három játékossal rendelkező lényeges játékok osztályára. □

Az 5.3., 5.4., 5.5. következmények nagyon hasonlóak a Shapley-féle megközelítésnél kapottakkal: rendre a 4.1., 4.2., 4.3. következményekkel. Tehát a két axiomatizáció meglehetősen hasonlít egymásra.

Megjegyzés. Ha még feltesszük az ETP tulajdonságot is a PO, NP és FP tulajdonságok megtartása mellett, akkor az 5.3. és 5.4. következmények továbbra is igazak maradnak, és az 5.5. következmény a következő képpen változik: „Ha $|N| = 2$, akkor van olyan a lényeges játékok osztályán értelmezett PO, NP, ETP és FP megoldás ψ, hogy $\psi \neq \phi$. Azonban, ha $|N| \neq 2$, akkor a lényeges játékok osztályán a PO, NP, ETP, FP axiómák jellemzik a Shapley-értéket."

6. Young axiomatizációja

Ebben a részben Young [22] axiomatizálását tárgyaljuk. A következő példa a rész fő eredményének – a 6.1. tételnek – gondolatát mutatja be.

6.1. Példa. Legyen $N \triangleq \{1,2,3\}$ és $v \triangleq (0,0,0,3,1,2,3)$. Ekkor v egy szuperadditív, de nem konvex játék, $v'_1 = (0,0,3,1,0,0,1)$, $v'_2 = (0,3,0,2,0,2,0)$, $v'_3 = (0,1,2,0,0,0,0)$ (az első komponens $v'_i(\emptyset)$ stb., az utolsó $v'_i(\{2,3\})$), így $1 \asymp^v 2$, $1 \asymp^v 3$ és $2 \asymp^v 3$.

Továbbá, legyen ψ egy a $\mathcal{G}^N$-n értelmezett PO, ETP és EMP megoldás. Azt mutatjuk meg, hogy $\psi_2(v) = \phi_2(v)$.

Rögzítsük az 1 játékost, és válasszuk a 2 játékost másolónak ($w'_2 = v'_2$ lásd később). Ekkor van olyan játék $w \triangleq (0,0,0,3,2,2,4)$[4], ahol $w'_1 = (0,0,3,2,0,0,2)$, $w'_2 = v'_2 = (0,3,0,2,0,2,0)$, $w'_3 = (0,2,2,0,1,0,0)$, hogy $1 \sim^w 2$ (ebben az esetben $1 \asymp^w 3$).

[4] Világos, hogy w nem az egyetlen játék, ahol $w'_2 = v'_2$ és $1 \sim^w 2$.

Most rögzítsük az $\{1,2\}$ halmazt, és legyen a 3 játékos a másoló. Ekkor van olyan játék $z \triangleq (0,0,0,2,2,2,3)$, ahol $z'_1 = (0,0,2,2,0,0,1)$, $z'_2 = (0,2,0,2,0,1,0)$, $z'_3 = w'_3 = (0,2,2,0,1,0,0)$, hogy $1 \sim^z 2 \sim^z 3$.

Ekkor PO és ETP következtében $\psi(z) = \phi(z)$. Továbbá, EMP miatt $\psi_3(w) = \phi_3(w)$. Mivel ψ PO és ETP, $1 \sim^w 2$, így $\psi(w) = \phi(w)$.

Megint alkalmazva EMP-t, azt kapjuk, hogy $\psi_2(v) = \phi_2(v)$.

Fontos látni, hogy a fenti példában tetszőleges i játékos esetén meg tudjuk mutatni, hogy $\psi_i = \phi_i$. Magyarán szólva $\psi(v) = \phi(v)$. Egyetlen dologra van csak szükségünk a levezetéshez, arra, hogy ψ legyen értelmezve a v-től a z-be vezető utak mentén (w és z függ a választott játékostól).

6.1. Definíció. Az $A \subseteq \mathcal{G}^N$ halmaz EMP-zárt, ha $\forall v \in A$-ra, hogy S ekvivalencia halmaz v-ben, és $\forall k \in N \setminus S$-re $\exists w \in A$, hogy $S \cup \{k\}$ ekvivalencia halmaz w-ben és $w'_k = v'_k$.

A következő tétel ennek a résznek a fő eredménye.

6.1. TÉTEL. *Legyen $A \subseteq \mathcal{G}^N$ olyan, hogy $\forall v \in A$-ra és $\forall k \in N$-re $\exists B \subseteq A$, $\exists w \in A$, és $\forall i \in N \setminus \{k\}$-ra $\exists z(i) \in B$, hogy*

1. *B EMP-zárt,*
2. *$w'_k = v'_k$ és $\forall i \in N \setminus \{k\}$-ra $z(i)'_i = w'_i$.*

Ekkor ψ az A-n értelmezett megoldás pontosan akkor PO, ETP és EMP, ha $\psi = \phi$.

Bizonyítás.

Szükséges: Lásd a 2.1. állítást.

Elégséges: Legyen $v \in A$ tetszőlegesen rögzített, és $n \triangleq |N|$. Továbbá, legyen $i_1 \in N$ tetszőlegesen rögzített és $i_2 \in N \setminus \{i_1\}$ szintén tetszőlegesen rögzített. Legyen B a tétel fejrészében meghatározott EMP-zárt halmaz (természetesen B függ v-től és i_1-től), továbbá legyen $z \in B$ tetszőlegesen rögzített.

Mivel B EMP-zárt, így $\exists z(1) \in B$, hogy $z(1)'_{i_2} = z'_{i_2}$ és $\{i_1, i_2\}$ ekvivalencia halmaz $z(1)$-ben. Legyen $i_3 \in N \setminus \{i_1, i_2\}$ tetszőlegesen rögzített.

Mivel B EMP-zárt, így $\exists z(2) \in B$, hogy $z(2)'_{i_3} = z(1)'_{i_3}$ és $\{i_1, i_2, i_3\}$ ekvivalencia halmaz $z(2)$-ben. Legyen $i_4 \in N \setminus \{i_1, i_2, i_3\}$ tetszőlegesen rögzített.

$\vdots$

Mivel B EMP-zárt, így $\exists z(n-1) \in B$, hogy $z(n-1)'_{i_n} = z(n-2)'_{i_n}$ és $\{i_1, i_2, \ldots, i_n\} = N$ ekvivalencia halmaz $z(n-1)$-ben.

ψ PO és ETP, továbbá értelmezve van B-n, így $\psi(z(n-1)) = \phi(z(n-1))$. Ekkor mivel ψ PO, ETP és EMP, és értelmezve van B-n, így

$$\psi(z(n-2)) = \phi(z(n-2)).$$

Mivel $i_{n-1} \in N \setminus \{i_1, \ldots, i_{n-2}\}$ tetszőlegesen rögzített volt, $\{i_1, \ldots, i_{n-2}\}$ ekvivalencia halmaz $z(n-3)$-ban, ψ PO, ETP és EMP, és értelmezve van B-n, így $\psi(z(n-3)) = \phi(z(n-3))$.

$\vdots$

Mivel $i_2 \in N \setminus \{i_1\}$ tetszőlegesen rögzített volt, ψ PO és EMP, és értelmezve van B-n, így $\psi(z) = \phi(z)$.

Legyen $k \triangleq i_1$, és w, $z(i)$ a tétel fejrészében meghatározottak. Mivel $\forall i \in N \setminus \{i_1\}$-re $w'_i = z(i)'_i$, $\forall z \in B$-re $\psi(z) = \phi(z)$, ψ PO és EMP, és értelmezve van A-n, így $\psi(w) = \phi(w)$.

$w'_{i_1} = v'_{i_1}$ és ψ EMP, tehát $\psi_{i_1}(v) = \phi_{i_1}(v)$.

i_1 tetszőlegesen rögzített volt, így $\psi(v) = \phi(v)$. □

A következőkben $\mathcal{G}^N$-t vizsgáljuk.

6.1. LEMMA. *Legyen $v \in \mathcal{G}^N$ tetszőlegesen rögzített. $S \subseteq N$ pontosan akkor ekvivalencia halmaz v-ben, ha $\forall T, Z \subseteq N$-re, hogy $T \setminus S = Z \setminus S$ és $|T| = |Z|$: $v(T) = v(Z)$.*

Bizonyítás. Szükséges: A bizonyítást az olvasóra hagyjuk.

Elégséges: Feltehetjük, hogy $T \setminus Z \neq \emptyset$, és legyen $m \triangleq |T \setminus Z| = |Z \setminus T|$. Mivel $T \setminus Z \subseteq S$ és $Z \setminus T \subseteq S$, S ekvivalencia halmaz v-ben, így

$$\begin{aligned}
v((T \cap Z) \cup \{l_1\}) &= v(T \cap Z) + v'_{l_1}(T \cap Z) = \\
&= v(T \cap Z) + v'_{q_1}(T \cap Z) = v((T \cap Z) \cup \{q_1\}), \\
&\qquad \text{ahol } l_1 \in T \setminus Z, \text{ és } q_1 \in Z \setminus T \\
v((T \cap Z) \cup \{l_1, l_2\}) &= v((T \cap Z) \cup \{l_1\}) + v'_{l_2}((T \cap Z) \cup \{l_1\}) = \\
&= v((T \cap Z) \cup \{q_1\}) + v'_{q_2}((T \cap Z) \cup \{q_1\}) = \\
&= v((T \cap Z) \cup \{q_1, q_2\}), \\
&\qquad \text{ahol } l_2 \in T \setminus \{Z \cup l_1\}, \text{ és } q_2 \in Z \setminus \{T \cup q_1\} \\
&\vdots \\
v((T \cap Z) \cup \{l_1, \ldots, l_m\}) &= v((T \cap Z) \cup \{l_1, \ldots, l_{m-1}\}) + \\
&\quad + v'_{l_m}((T \cap Z) \cup \{l_1, \ldots, l_{m-1}\}) = \\
&= v((T \cap Z) \cup \{q_1, \ldots, q_{m-1}\}) + \\
&\quad + v'_{q_m}((T \cap Z) \cup \{q_1, \ldots, q_{m-1}\}) = \\
&= v((T \cap Z) \cup \{q_1, \ldots, q_m\}), \\
&\qquad \text{ahol } l_m \in T \setminus \{Z \cup \{l_1, \ldots, l_{m-1}\}, \\
&\qquad \text{és } q_m \in Z \setminus \{T \cup \{q_1, \ldots, q_{m-1}\} \\
v(T) &= v((T \cap Z) \cup \{l_1, \ldots, l_m\}) = \\
&= v((T \cap Z) \cup \{q_1, \ldots, q_m\}) = v(Z).
\end{aligned}$$

□

A 6.1. segédtételből közvetlenül következik a következő állítás.

6.1. KÖVETKEZMÉNY. *Legyen $v \in \mathcal{G}^N$ tetszőlegesen rögzített, $S \subset N$ ekvivalencia halmaz v-ben, és $k \in N \setminus S$ szintén tetszőlegesen rögzített. Ekkor $\forall T, Z \subseteq N$-re, hogy $T \setminus S = Z \setminus S$ és $|T| = |Z|$: $v'_k(T) = v'_k(Z)$.*

6.2. LEMMA. *$\mathcal{G}^N$ EMP-zárt.*

Bizonyítás. Legyen $v \in \mathcal{G}^N$ olyan, hogy $S \subset N$ ekvivalencia halmaz v-ben, és $k \in N \setminus S$ tetszőlegesen rögzített.

Ha $T = \emptyset$, akkor legyen $w(T) \triangleq 0$. Ha $T \cap (S \cup \{k\}) = \emptyset$, $T \neq \emptyset$, akkor legyen $w(T)$ tetszőlegesen rögzített. Különben $(T \cap (S \cup \{k\}) \neq \emptyset)$, legyen

$$w(T) \triangleq w(T \setminus (S \cup \{k\})) + \sum_{i=1}^{m} v'_k((T \setminus (S \cup \{k\})) \cup \{l_1, \ldots, l_{i-1}\}), \qquad (9)$$

ahol $m \triangleq |(S \cup \{k\}) \cap T|$, és $l_i \in S \cap T$, $i = 1, \ldots, m-1$. Vegyük észre, hogy a 6.1. következmény miatt

$$\sum_{i=1}^{m} v'_k((T \setminus (S \cup \{k\})) \cup \{l_1, \ldots, l_{i-1}\})$$

nem függ $S \cap T$ elemeinek sorrendjétől.

Világos, hogy $w'_k = v'_k$, továbbá, a 6.1. segédtételből $S \cup \{k\}$ ekvivalencia halmaz w-ben. □

A következő segédtétel bizonyítása az Appendixben található.

6.3. LEMMA. *A szigorúan konvex / additív / szigorúan konkáv játékok osztálya EMP-zárt.*

A következő segédtétel a 6.1. tétel 2. pontjához kötődik.

6.4. LEMMA. *Nézzük a következő pontokat:*

1. *Legyen $v \in \mathcal{G}^N$ tetszőlegesen rögzített lényeges / konvex / (szigorúan) szuperadditív / (szigorúan) gyengén-szuperadditív / (szigorúan) monoton játék, és $k \in N$ tetszőlegesen rögzített. Ekkor $\exists w$ lényeges / konvex / (szigorúan) szuperadditív / (szigorúan) gyengén-szuperadditív / (szigorúan) monoton játék, és $\forall i \in N \setminus \{k\}$-hoz $\exists z(i)$ olyan szigorúan konvex játék, hogy $w'_k = v'_k$ és $\forall i \in N \setminus \{k\}$-ra $z(i)'_i = w'_i$.*
2. *Legyen $v \in \mathcal{G}^N$ tetszőlegesen rögzített (szigorúan) gyengén-szubadditív / (szigorúan) szubadditív / konkáv játék, és $k \in N$ tetszőlegesen rögzített. Ekkor $\exists w$ (szigorúan) gyengén-szubadditív / (szigorúan) szubadditív / konkáv játék, és $\forall i \in N \setminus \{k\}$-hoz $\exists z(i)$ olyan szigorúan konkáv játék, hogy $w'_k = v'_k$ és $\forall i \in N \setminus \{k\}$-ra $z(i)'_i = w'_i$.*

Bizonyítás. Az 1. pont: Legyen $M > \max\limits_{T \subset N} \mid v'_k(T) \mid$ tetszőlegesen rögzített. Továbbá, legyen

$$w(T) \triangleq 2M|N|3^{|T|},$$

ahol T olyan, hogy $k \notin T$, $T \neq \emptyset$, és legyen $w(\emptyset) \triangleq 0$.

Ha $k \in T$, akkor legyen $w(T) \triangleq w(T \setminus \{k\}) + v'_k(T \setminus \{k\})$. Könnyen látható, hogy w lényeges / konvex / (szigorúan) szuperadditív / (szigorúan) gyengén-szuperadditív / (szigorúan) monoton játék és $w'_k = v'_k$. Továbbá, legyen $l \in N \setminus \{k\}$, és $T, Z \subseteq N \setminus \{l\}$, $Z \subset T$ tetszőlegesen rögzített. Ekkor három eset lehetséges.

$k \in Z$: Ekkor

$$\begin{aligned} w'_l(T) - w'_l(Z) &= w(T \cup \{l\}) - w(T) - w(Z \cup l) + w(Z) = \\ &= w((T \setminus \{k\}) \cup \{l\}) + w'_k((T \setminus \{k\}) \cup \{l\}) - w(T \setminus \{k\}) - w'_k(T \setminus \{k\}) - \\ &- w((Z \setminus \{k\}) \cup \{l\}) - w'_k((Z \setminus \{k\}) \cup \{l\}) + w(Z \setminus \{k\}) + w'_k(Z \setminus \{k\}). \end{aligned}$$

Továbbá,

$$w'_k((T \setminus \{k\}) \cup \{l\}) - w'_k(T \setminus \{k\}) > -2M,$$

és

$$w'_k(Z \setminus \{k\}) - w'_k((Z \setminus \{k\}) \cup \{l\}) > -2M.$$

Összefoglalva a fenti egyenlőtlenségeket

$$\begin{aligned} &w'_k((T \setminus \{k\}) \cup \{l\}) - w'_k(T \setminus \{k\}) - \\ &- w'_k((Z \setminus \{k\}) \cup \{l\}) + w'_k(Z \setminus \{k\}) > -4M. \end{aligned} \tag{10}$$

Továbbá,

$$w((T \setminus \{k\}) \cup \{l\}) - w(T \setminus \{k\}) = 4M|N|3^{|T|-1},$$

és

$$w((Z \setminus \{k\}) \cup \{l\}) - w(Z \setminus \{k\}) = 4M|N|3^{|Z|-1}.$$

Tehát

$$\begin{aligned} w((T \setminus \{k\}) \cup \{l\}) - w(T \setminus \{k\}) - w((Z \setminus \{k\}) \cup \{l\}) + w(Z \setminus \{k\}) = \\ = 4M|N|3^{|Z|-1}(3^{|T \setminus Z|} - 1). \end{aligned} \tag{11}$$

Összefoglalva a (10) és (11) egyenlőtlenségeket

$$w'_l(T) - w'_l(Z) > 0.$$

A másik két eset bizonyítását ($k \notin T$ és $k \in T \setminus Z$) az olvasóra bízzuk.

Legyen $i \in N \setminus \{k\}$ tetszőlegesen rögzített. Megismételve a fenti eljárást ($v = w$, $k = i$) megkapjuk $z(i)$-t. Ekkor $z(i)'_i = w'_i$, és $\forall T, Z \subseteq N \setminus \{i\}$-re, hogy $Z \subset T$: $w'_i(T) - w'_i(Z) > 0$, tehát $z(i)$ szigorúan konvex játék.

A 2. pont: A (szigorúan) gyengén-szubadditív / (szigorúan) szubadditív / konkáv játékok osztálya tartalmazza a szigorúan konkáv játékok osztályát. Könnyen látható, hogy vehetjük tetszőleges (szigorúan) gyengén-szubadditív / (szigorúan) szubadditív / konkáv játék duálisát, és alkalmazhatjuk az 1. pontot[5], majd vehetjük az 1. pont által produkált játékok duálisait. □

A 6.4. segédtétel azért fontos, mert nem minden vizsgált játékosztály EMP-zárt.

Megjegyzés. A 6.3. segédtétel bizonyításából látható, hogy a (szigorúan) konvex, (szigorúan) gyengén-szuperadditív, (szigorúan) monoton, additív, (szigorúan) gyengén-szubadditív, (szigorúan) konkáv játékosztályok EMP-zártak. Az a közös ezekben a játékosztályokban, hogy jól jellemezhetőek játékosaik határhozzájárulási függvényeivel. Ez a tulajdonság felelős az EMP-zártságért.

A lényeges, (szigorúan) szuperadditív, (szigorúan) szubadditív játékosztályok azonban nem EMP-zártak.

6.2. Példa. (1) Legyen $v \triangleq (0, 0, 10, 50, 0, 0, 20)$, ahol $S \triangleq \{1, 2\}$ ekvivalencia halmaz v-ben. v lényeges játék, azonban az egyetlen olyan játék, amiben N ekvivalencia halmaz és $w'_3 = v'_3$ a $(10, 10, 10, 10, 10, 10, -20)$, ami nem lényeges játék.

(2) Legyen $v \triangleq (0, 0, 0, 10, 51, 51, 51, 51, 51, 51, 62, 62, 62, 62, 103)$, ahol $S \triangleq \{1, 2, 3\}$ ekvivalencia halmaz v-ben. v szigorúan szuperadditív játék, de az egyetlen olyan játék amelyben N ekvivalencia halmaz és $w'_4 = v'_4$ a $(10, 10, 10, 10, 61, 61, 61, 61, 61, 61, 72, 72, 72, 72, 113)$, ami nem szuperadditív játék.

(3) Legyen $v \triangleq (100, 100, 100, 10, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0)$, ahol $S \triangleq \{1, 2, 3\}$ ekvivalencia halmaz v-ben. v szigorúan szubadditív játék, de az egyetlen olyan játék amelyben N ekvivalencia halmaz és $w'_4 = v'_4$ a $(10, 10, 10, 10, -89, -89, -89, -89, -89, -89, -90, -90, -90, -90, -90)$, ami nem szubadditív játék.

Megjegyzés. Ha $|N| \leq 3$, akkor a (szigorúan) szuperadditív, (szigorúan) szubadditív játékosztályok (N a játékosok halmaza) rendre egybeesnek a (szigorúan) gyengén-szuperadditív, (szigorúan) gyengén-szubadditív játékosztályokkal, így EMP-zártak. Továbbá, ha $|N| \leq 2$, akkor a lényeges játékok osztálya egybeesik a szigorúan konvex játékok osztályával, így EMP-zárt.

A fenti eredményeket (6.1. tétel, 6.3., 6.4. segédtételek) összefoglalva a következő eredményt kapjuk.

[5]Fontos látni, hogy tetszőleges (szigorúan) szubadditív / (szigorúan) gyengén-szubadditív játék duálisa nem feltétlenül (szigorúan) szuperadditív / (szigorúan) gyengén-szuperadditív játék. pl. $v \triangleq (4, 4, 4, 4, 4, 4, 7)$ szigorúan szubadditív, de $\bar{v}$ nem gyengén-szuperadditív. Továbbá, tetszőleges (szigorúan) szuperadditív / (szigorúan) gyengén-szuperadditív játék duálisa nem feltétlenül (szigorúan) szubadditív / (szigorúan) gyengén-szubadditív játék. Pl. $v \triangleq (0, 0, 0, 3, 1, 2, 4)$ szigorúan szuperadditív, de $\bar{v}$ nem gyengén-szubadditív.

6.2. KÖVETKEZMÉNY. *ψ a lényeges / (szigorúan) konvex / (szigorúan) szuperadditív / (szigorúan) gyengén-szuperadditív / (szigorúan) monoton / additív / (szigorúan) gyengén-szubadditív / (szigorúan) szubadditív / (szigorúan) konkáv játékok osztályán értelmezett megoldás pontosan akkor PO, ETP és EMP, ha $\psi = \phi$.*

7. Összefoglalás

A 3.1., 3.2., 4.1., 4.2., 4.3., 5.3., 5.4., 5.5., 6.2. következmények és a 4.2. következményt követő megjegyzés összefoglalása az 1. táblázatban látható.

$\surd$ azt jelenti, hogy az adott karakterizáció (oszlop) érvényes az adott játékosztályon (sor). Zárójelek között arra utalunk, aki először kapta meg az adott eredményt. $\emptyset$ azt jelenti, hogy az adott karakterizáció (oszlop) nem érvényes az adott játékosztályon (sor). Végül, szögletes zárójelek között a feltételt adjuk meg, amely mellet az adott jellemzés igaz, pl. $\surd$ $[|N| \neq 2]$ azt jelenti, hogy az adott játékosztályon ha $|N| \neq 2$, akkor érvényes, ha $|N| = 2$, akkor nem érvényes az adott karakterizáció.

8. Appendix

Bizonyítás.[A 6.3. segédtétel bizonyítása] Legyen $v \in \mathcal{G}^N$ olyan, hogy $S \subset N$ ekvivalencia halmaz v-ben, és $k \in N \setminus S$ tetszőlegesen rögzített. A 6.2. segédtétel bizonyításából látszik, hogy $\exists w \in \mathcal{G}^N$, hogy $S \cup \{k\}$ ekvivalencia halmaz w-ben és $w'_k = v'_k$. Továbbá, $\forall T \subseteq N$-re, hogy $T \cap (S \cup \{k\}) = \emptyset$, $T \neq \emptyset$: $w(T)$ tetszőlegesen rögzített lehet. Tehát az egyetlen dolog, amit meg kell mutatnunk (kivéve az additív játékok triviális esetét), hogy tudunk olyan értékeket rendelni ezekhez a koalíciókhoz, hogy az így kapott w benne legyen a kívánt játékosztályban.

(1) Az additív játékok osztálya: Köztudott, hogy $z \in \mathcal{G}^N$ pontosan akkor additív, ha $\forall i \in N$-re $\exists c_i \in \mathbb{R}$, hogy $\forall T \subseteq N \setminus \{i\}$-re $z'_i(T) = c_i$.
Legyen $c^* \triangleq v'_k(\emptyset)$. Továbbá, $\forall T \subseteq N$-re legyen

$$w(T) \triangleq c^*|T|.$$

Világos $w'_k = v'_k$, w additív, és N ekvivalencia halmaz w-ben.

(2) A szigorúan konvex játékok osztálya: A $z \in \mathcal{G}^N$ játék pontosan akkor szigorúan konvex, ha $\forall i \in N$-re, $\forall T, Z \subseteq N \setminus \{i\}$-re, hogy $Z \subset T$: $z'_i(Z) < z'_i(T)$ (lásd a 2.1. segédtételt).
Legyen $M > \max\limits_{T \subset N} | v'_k(T) |$ tetszőlegesen rögzített. Továbbá, legyen

$$w(T) \triangleq M|N|3^{|T|}, \tag{12}$$

ahol $T \cap (S \cup \{k\}) = \emptyset$, $T \neq \emptyset$, és legyen $w(\emptyset) \triangleq 0$.

	Hart és Mas-Colell [8]	Shapley [20]	van den Brink [1]	Young [22]
lényeges	$\emptyset$	$\surd$ [$\lvert N\rvert \neq 2$]	$\surd$ [$\lvert N\rvert \neq 2,3$]	$\surd$
szigorúan konvex	$\surd$	$\emptyset$ [$\lvert N\rvert > 1$]	$\emptyset$ [$\lvert N\rvert > 1$]	$\surd$
konvex	$\surd$	$\surd$ ([16])	$\surd$	$\surd$
szigorúan szuperadditív	$\surd$	$\emptyset$ [$\lvert N\rvert > 1$]	$\emptyset$ [$\lvert N\rvert > 1$]	$\surd$
szuperadditív	$\surd$	$\surd$ ([20])	$\surd$	$\surd$ ([22])
szigorúan gyengén-szuperadditív	$\surd$	$\emptyset$ [$\lvert N\rvert > 1$]	$\emptyset$ [$\lvert N\rvert > 1$]	$\surd$
gyengén-szuperadditív	$\surd$	$\surd$ ([16])	$\surd$	$\surd$
szigorúan monoton	$\surd$	$\emptyset$ [$\lvert N\rvert > 1$]	$\emptyset$ [$\lvert N\rvert > 1$]	$\surd$
monoton	$\surd$	$\surd$ ([7])	$\surd$	$\surd$
additív	$\surd$	$\surd$ ([16])	$\surd$	$\surd$
gyengén-szubadditív	$\surd$	$\surd$	$\surd$	$\surd$
szigorúan gyengén-szubadditív	$\surd$	$\emptyset$ [$\lvert N\rvert > 1$]	$\emptyset$ [$\lvert N\rvert > 1$]	$\surd$
szubadditív	$\surd$	$\surd$	$\surd$	$\surd$
szigorúan szubadditív	$\surd$	$\emptyset$ [$\lvert N\rvert > 1$]	$\emptyset$ [$\lvert N\rvert > 1$]	$\surd$
konkáv	$\surd$	$\surd$	$\surd$	$\surd$
szigorúan konkáv	$\surd$	$\emptyset$ [$\lvert N\rvert > 1$]	$\emptyset$ [$\lvert N\rvert > 1$]	$\surd$

1. táblázat. A Shapley-érték axiomatizációi

Legyen $l \in N \setminus (S \cup \{k\})$ tetszőlegesen rögzített, és $T, Z \subseteq N \setminus \{l\}$ olyan, hogy $Z \subset T$. Ekkor (9)-ből

$$\begin{aligned} w'_l(T) &= w(T \cup \{l\}) - w(T) = \\ &= w((T \cup \{l\}) \setminus (S \cup \{k\})) + \\ &+ \sum_{i=1}^{m} w'_{l_i}(((T \cup \{l\}) \setminus (S \cup \{k\})) \cup \{l_1, \ldots, l_{i-1}\}) - \\ &- w(T \setminus (S \cup \{k\})) - \sum_{i=1}^{m} w'_{l_i}((T \setminus (S \cup \{k\})) \cup \{l_1, \ldots, l_{i-1}\}), \end{aligned}$$

és

$$\begin{aligned} w'_l(Z) &= w(Z \cup \{l\}) - w(Z) = \\ &= w((Z \cup \{l\}) \setminus (S \cup \{k\})) + \\ &+ \sum_{i=1}^{n} w'_{l_i}(((Z \cup \{l\}) \setminus (S \cup \{k\})) \cup \{l_1, \ldots, l_{i-1}\}) - \\ &- w(Z \setminus (S \cup \{k\})) - \sum_{i=1}^{n} w'_{l_i}((Z \setminus (S \cup \{k\})) \cup \{l_1, \ldots, l_{i-1}\}), \end{aligned}$$

ahol $m \triangleq |(S \cup \{k\}) \cap T|$, $n \triangleq |(S \cup \{k\}) \cap Z|$, és

$$\begin{aligned} \{l_1, \ldots, l_n\} &\triangleq (S \cup \{k\}) \cap Z = (S \cup \{k\}) \cap (Z \cup \{l\}) \subseteq \{l_1, \ldots, l_m\} \triangleq \\ &\triangleq (S \cup \{k\}) \cap T = (S \cup \{k\}) \cap (T \cup \{l\}). \end{aligned}$$

Vegyük észre, hogy ha $T \setminus (S \cup \{k\}) = Z \setminus (S \cup \{k\})$, akkor a bizonyítás kész. Tegyük tehát fel, hogy $Z \setminus (S \cup \{k\}) \subset T \setminus (S \cup \{k\})$. Mivel v szigorúan konvex játék és $S \cup \{k\}$ ekvivalencia halmaz w-ben, így $\forall i \leq n$-re

$$\begin{aligned} 2M > w'_{l_i}(((T \cup \{l\}) \setminus (S \cup \{k\})) \cup \{l_1, \ldots, l_{i-1}\}) - \\ - w'_{l_i}((T \setminus (S \cup \{k\})) \cup \{l_1, \ldots, l_{i-1}\}) > 0, \end{aligned}$$

és

$$\begin{aligned} 2M > w'_{l_i}(((Z \cup \{l\}) \setminus (S \cup \{k\})) \cup \{l_1, \ldots, l_{i-1}\}) - \\ - w'_{l_i}((Z \setminus (S \cup \{k\})) \cup \{l_1, \ldots, l_{i-1}\}) > 0, \end{aligned}$$

így

$$\begin{aligned} w'_{l_i}(((T \cup \{l\}) \setminus (S \cup \{k\})) \cup \{l_1, \ldots, l_{i-1}\}) - \\ - w'_{l_i}((T \setminus (S \cup \{k\})) \cup \{l_1, \ldots, l_{i-1}\}) - \\ - w'_{l_i}(((Z \cup \{l\}) \setminus (S \cup \{k\})) \cup \{l_1, \ldots, l_{i-1}\}) + \\ + w'_{l_i}((Z \setminus (S \cup \{k\})) \cup \{l_1, \ldots, l_{i-1}\}) > -2M. \end{aligned}$$

$n < |N|$, tehát

$$\begin{aligned}&\sum_{i=1}^{m} w'_{l_i}(((T\cup\{l\})\setminus(S\cup\{k\}))\cup\{l_1,\dots,l_{i-1}\}) -\\ &-\sum_{i=1}^{m} w'_{l_i}((T\setminus(S\cup\{k\}))\cup\{l_1,\dots,l_{i-1}\}) -\\ &-\sum_{i=1}^{n} w'_{l_i}(((Z\cup\{l\})\setminus(S\cup\{k\}))\cup\{l_1,\dots,l_{i-1}\}) +\\ &+\sum_{i=1}^{n} w'_{l_i}((Z\setminus(S\cup\{k\}))\cup\{l_1,\dots,l_{i-1}\}) > -2M|N|.\end{aligned} \tag{13}$$

Ugyanakkor, (12)-ből

$$w((T\cup\{l\})\setminus(S\cup\{k\})) - w(T\setminus(S\cup\{k\})) = 2M|N|3^{|T\setminus(S\cup\{k\})|},$$

és

$$w((Z\cup\{l\})\setminus(S\cup\{k\})) - w(Z\setminus(S\cup\{k\})) = 2M|N|3^{|Z\setminus(S\cup\{k\})|},$$

így $Z\subset T$-ből következik, hogy (emlékezzünk $Z\setminus(S\cup\{k\})\subset T\setminus(S\cup\{k\})$)

$$\begin{aligned}&w((T\cup\{l\})\setminus(S\cup\{k\})) - w(T\setminus(S\cup\{k\})) -\\ &-w((Z\cup\{l\})\setminus(S\cup\{k\})) + w(Z\setminus(S\cup\{k\})) =\\ &=2M|N|3^{|Z\setminus(S\cup\{k\})|}(3^{|(T\setminus Z)\setminus(S\cup\{k\})|}-1) > 2M|N|.\end{aligned} \tag{14}$$

Összefoglalva a (13) és (14) egyenlőtlenségeket

$$w'_l(T) - w'_l(Z) > 0.$$

$w'_k = v'_k$, v szigorúan konvex, $\forall i\in S\cup\{k\}$-re $i\sim^w k$, továbbá $l\in N\setminus(S\cup\{k\})$ és $T, Z\subseteq N\setminus\{l\}$, $Z\subset T$ tetszőlegesen rögzítettek voltak, így w szigorúan konvex játék.

(3) A szigorúan konkáv játékok osztálya: A $z\in\mathcal{G}^N$ játék pontosan akkor szigorúan konkáv, ha $\forall i\in N$-re, $\forall T, Z\subseteq N\setminus\{i\}$-re, hogy $Z\subset T$: $z'_i(Z) > z'_i(T)$ (lsd. a 2.1. segédtételt).

A 2.2. segédtételből $\bar{v}$ szigorúan konvex játék és S ekvivalencia halmaz $\bar{v}$-ben. Ekkor a (2) pontból $\exists z$ olyan szigorúan konvex játék, hogy $S\cup\{k\}$ ekvivalencia halmaz z-ben, és $z'_k = \bar{v}'_k$. A 2.2. segédtétel miatt $\bar{z}$ szigorúan konkáv játék, és $S\cup\{k\}$ ekvivalencia halmaz z-ben.

Megmutatjuk, hogy $\bar{z}'_k = v'_k$.

$\forall T\subseteq N\setminus\{k\}$-ra

$$\bar{v}'_k(T) = v(N\setminus T) - v(N\setminus(T\cup\{k\})) = v'_k(N\setminus(T\cup\{k\})),$$

így $\forall T \subseteq N \setminus \{k\}$-ra

$$z'_k(T) = v'_k(N \setminus (T \cup \{k\})).$$

Magyarán szólva, $\forall T \subseteq N \setminus \{k\}$-ra

$$z'_k(N \setminus (T \cup \{k\})) = v'_k(N \setminus ((N \setminus (T \cup \{k\})) \cup \{k\})) = v'_k(T).$$

z duálisát véve, $\forall T \subseteq N \setminus \{k\}$-ra

$$\bar{z}'_k(T) = z'_k(N \setminus (T \cup \{k\})),$$

tehát $\forall T \subseteq N$-re

$$\bar{z}'_k(T) = v'_k(T).$$

Végül, legyen $w \triangleq \bar{z}$.

□

Hivatkozások

[1] VAN DEN BRINK R.: *An axiomatization of the Shapley value using a fairness property*, International Journal of Game Theory **30**, 309–319. (2001)

[2] CHUN Y.: *A New Axiomatization of the Shapley Value*, Games and Economic Behavior **1**, 119–130. (1989)

[3] CHUN Y.: *On the Symmetric and Weighted Shapley Values*, International Journal of Game Theory **20**, 183–190. (1991)

[4] CSÓKA P.: *Koherens kockázatmérés és tőkeallokáció*, Közgazdasági Szemle **50**, 855–880. (2003)

[5] DRIESSEN T.: *Cooperative games, sollutions and applications*, Kluwer Academic Press, Boston / Dordrecht / London (1988)

[6] DUBEY P.: *On the uniqueness of the Shapley value*, International Journal of Game Theory **4**, 131–139. (1975)

[7] EINY E.: *The Shapley value on some lattices of monotonic games*, Mathematical Social Sciences **15**, 1–10. (1988)

[8] HART S., MAS-COLELL A.: *Potential, value, and consistency*, Econometrica **57**, 589–614. (1989)

[9] HART S., A. MAS-COLELL: *The Potential of the Shapley Value*, The Shapley Value, Roth A. E. (szerkesztő), Cambridge University Press, 127–137. (1988)

[10] LANGE F., M. GRABISCH: *A recursive solution concept for multichoice games*, Acta Polytechnica Hungarica **5**, 47–57. (2008)

[11] MORETTI S., F. PATRONE: *Transversality of the Shapley value*, Top **16**, 1–41. (2008)

[12] MOULIN H.: *Axioms of cooperative decision making*, Cambridge University Press (1988)

[13] MYERSON R. B.: *Graphs and cooperation in games*, Mathematics of Operations Research **2**, 225–229. (1977)

[14] NEYMAN A.: *Uniqueness of the Shapley value*, Games and Economic Behavior **1**, 116–118. (1989)

[15] VAN DEN NOUWELAND A., TIJS S., MASCHLER M.: *Monotonic games are spanning network games*, International Journal of Game Theory **21**, 419–427. (1993)

[16] PELEG B., SUDHÖLTER P.: *Introduction to the Theory of Cooperative Games*, Kluwer Academic Publishers, Boston / Dordrecht / London (2003)

[17] PINTÉR M.: *Regressziós játékok*, Szigma **38**, 131–148. (2007)

[18] ROTH A. E.: *The Shapley Value as a von Neumann-Morgenstern Utility*, Econometrica **45**, 657–664. (1977)

[19] ROTH A. E.: *The expected value of playing a game* in A. E. Roth, ed., The Shapley Value, Cambridge University Press, 51–70. (1988)

[20] SHAPLEY L. S.: *A Value for n-Person Games*, Contributions to the Theory of Games Volume II (Annals of Mathematical Studies **28**, editors: Kuhn, H. W. – Tucker, A. W.) 307–317. (1953)

[21] SHAPLEY L. S.: *A Comparison of Power Indices and a Nonsymmetric Generalization*, P-5872, The Rand Corporation, Santa Monica, CA. (1977)

[22] YOUNG H. P.: *Monotonic Solutions of Cooperative Games*, International Journal of Game Theory **14**, 65–72. (1985)

(Beérkezett: 2008. március 19.)

PINTÉR MIKLÓS
Budapesti Corvinus Egyetem, Matematika Tanszék
1093 Budapest, Fővám tér 13–15.
miklos.pinter@uni-corvinus.hu

ON AXIOMATIZATIONS OF THE SHAPLEY VALUE[6]

MIKLÓS PINTÉR

The Shapley value is one of the most popular solution concepts for games in coalitional form. It is usual in the literature to axiomatize the Shapley value. In this paper we consider four axiomatizations of the Shapley value: Hart and Mas-Colell's approach based on potential, Shapley's original, van den Brink's and Young's characterizations. We examine the validity of the above four characterizations on sixteen sub-classes of transferable utility games. We summarize our results in a table.

[6]The author thanks the Hungarian Scientific Research Fund (OTKA) and the János Bolyai Research Scholarship of the Hungarian Academy of Sciences for financial support.

Alkalmazott Matematikai Lapok **26** (2009), 317-328.

VÉGTELEN DIFFERENCIÁLEGYENLET-RENDSZEREK STABILITÁSA

LUKIC ANIKÓ[1]

Az alábbi dolgozat végtelen lineáris egyenletrendszerek globális stabilitásának vizsgálatával foglalkozik minimális stabilizáló mechanizmus esetében. Az egyenletrendszer megoldóképletének felírása a Feynman–Kac-formulán alapszik. A megoldóképlet kiértékelése a Markov-láncok elméletének alkalmazásával történik. A cikk célja annak bemutatása, hogy ha egy adott végtelen összefüggőgráfon végbemenő Markov-lánc rekurrens, akkor a gráfon definiált egyenletrendszer stabil. A dolgozat végén a tranziens eset is tárgyalásra kerül.

1. Bevezető

A differenciálegyenletek elméletének gyakorlatban történő alkalmazása során gyakran adódnak olyan $\dot{x_k} = f(x_1, \ldots, x_n)$, $k = 1, 2, \ldots, n$ alakú egyenletrendszerek, ahol a jobboldali függvény csak az $x_k - x_j$ különbségektől függ. Jól látható, hogy az ilyen esetekben minden konstans konfiguráció $x = (x_1, x_2, \ldots, x_n)$, $x_i = c$, $i = 1, 2, \ldots, n$, $c \in \mathbb{R}$, egyben az egyenletrendszer stacionárius pontja is. Mivel azonban a stacionárius pontok halmaza összefüggő és kontinuum számosságú, az ilyen egyenletekből álló rendszer általában nem stabil. Ebben a dolgozatban egy egyszerű feladatot tárgyalunk, ahol megmutatjuk, hogy mindössze egyetlen egyenlet módosításával a rendszer stabilissá tehető. E kérdéskör különösen akkor válik érdekessé, ha a rendszer mérete a végtelenhez tart. A feladat a következőképpen fogalmazható meg:

Adott egy $\mathcal{G} = (G, E)$ véges vagy megszámlálhatóan végtelen összefüggőgráf. G_k jelöli a k csúcs szomszédos csúcsainak halmazát, mindig feltesszük, hogy $\sup |G_k| < \infty$. A $\mathcal{G}$ gráfon adott a következőegyenletrendszer:

$$\dot{x_k} = -\frac{\partial H}{\partial x_k}, \quad k \in G,$$

ahol $x_k \in \mathbb{R}$ és

$$H = \frac{1}{2}\Gamma(x_0) + \frac{1}{2}\sum_{k \in G}\sum_{j \in G_k} V(x_k - x_j).$$

[1]Támogatta: OTKA TS 49835 pályázat

Ekkor a H-hoz tartozó egyenletrendszer a következő

$$x_k = -\sum_{j \in G_k} V'(x_k - x_j), \quad k \neq 0, \quad \text{és} \tag{1}$$
$$x_0 = -\frac{1}{2}\Gamma'(x_0) - \sum_{j \in G_0} V'(x_0 - x_j).$$

Megjegyzés. Az (1) rendszerből jól látható, hogy $\Gamma = 0$ esetében minden konstans konfiguráció egyben a rendszer stacionárius pontja is, ami nem feltételenül következik, ha $\Gamma \neq 0$.

A következő fejezetben véges gráfokon definiált rendszerek exponencionális stabilizálhatóságát szemléltetjük *Ljapunov* módszere segítségével. A későbbiekben megmutatjuk, hogy a Markov-folyamatok elméletével, nevezetesen a Feynman–Kac-formula segítségével ilyen stabilitási feltételek, a Γ és V függvényekre vett természetes feltételek mellett, végtelen gráfokra is bizonyíthatóak, de a konvergencia sebessége ilyenkor már nem exponencionális. A tárgyalás folyamatossága érdekében néhány ismert, de nem közismert tény bizonyítását is közöljük.

2. Véges rendszerek

Ebben a fejezetben *Ljapunov* módszere segítségével röviden szemléltetjük, hogy a fenti típusú véges rendszerek exponenciálisan stabilak. Feltételek garantálják az egyértelmű megoldás létezését, ami a bizonyításból is látható.

2.1. TÉTEL. *Adott az (1) egyenletrendszer, ahol* $|G| < \infty$. *Legyen* $V \in C^2(\mathbb{R})$ *szimmetrikus függvény,* $V''(x), \Gamma''(x) \geq \alpha > 0 \quad \forall, x \in \mathbb{R}$, *valamint* $\Gamma'(0) = 0$. *Ekkor az (1) véges egyenletrendszer globálisan exponencionálisan stabil.*

Bizonyítás. Tekintsük a

$$Q(t) = \sum_{k \in G} x_k^2(t), \quad Q(0) < \infty,$$

kvadratikus *Ljapunov*-függvényt és annak a t szerinti deriváltját

$$\dot{Q}(t) = 2\sum_{k \in G} x_k \dot{x_k} = -x_0\Gamma'(x_0) - \sum_{k \in G} x_k \sum_{j \in G_k} V'(x_k - x_j).$$

Átrendezés után adódik, hogy

$$\dot{Q}(t) = -x_0\Gamma'(x_0) - \sum_{(k,j) \in E} (x_k - x_j)V'(x_k - x_j),$$

ahonnan jól látható, hogy a 0 konfiguráció a rendszer egyetlen stacionárius pontja.

A V és Γ függvények konvexitása és $V'(0) = \Gamma'(0) = 0$ miatt bármely $y \in \mathbb{R}$ esetén $yV'(y) = y^2V''(\xi) \geq \alpha y^2$, és ugyanígy, $y\Gamma'(y) \geq \alpha y^2$, tehát

$$\dot{Q}(t) \leq -\alpha x_0^2 - \alpha \sum_{(k,j)\in E} (x_k - x_j)^2.$$

Ezután a Cauchy-egyenlőtlenséggel kapjuk, hogy

$$x_k^2 = \Big(x_0 + (x_1 - x_0) + \cdots + (x_k - x_{k-1})\Big)^2 \leq$$

$$\leq (k+1)\Big(x_0^2 + (x_1 - x_0)^2 + \cdots + (x_k - x_{k-1})^2\Big) \leq -\frac{N\dot{Q}(t)}{\alpha},$$

ahol N az x_0 középpontú gráf sugara. A kapott eredményt összegezve

$$Q(t) = \sum_{k\in G} x_k^2 \leq -\sum_{k\in G} \frac{N\dot{Q}(t)}{\alpha} = -\frac{N|G|\dot{Q}(t)}{\alpha},$$

tehát $\dot{Q}(t) \leq -\sigma Q(t)$, ahol $\sigma := \frac{\alpha}{N|G|}$. Grönwall lemmájával $Q(t) \leq Q(0)e^{-\sigma t}$, vagyis a kívánt globális exponenciális stabilitás érvényes. □

Végezetül vegyük észre, hogy ha a rendszer mérete $|G| \to \infty$, akkor $\sigma \to 0$, melynek következtében végtelen rendszerek esetében az imént alkalmazott módszer nem vezet eredményhez. A dolgozat további részében a fenti egyenletrendszer lineáris változatának a végtelenbe történőkiterjesztésével foglalkozunk. A megoldó képlet felírása, illetve annak kiértéklése a Feynman–Kac-formula és a Markov-folyamatok elméletének alkalmazásával történik.

3. Bolyongás megszámlálható halmazon

Adott egy tetszőleges $\mathcal{G} = (G, E)$ összefüggőgráf, ahol G véges vagy megszámlálhatóan végtelen halmaz. Célunk a $\mathcal{G}$ gráfon olyan folytonos idejű bolyongást definiálni, ami minden más értelemben megfelelel a G állapotterű diszkrét idejű Markov-láncnak. Erre azért van szükség, mert a későbbi vizsgálódásaink során e folytonos idejű Markov-lánc rekurrencia tulajdonságára lesz szükségünk, viszont a rekurrenciára vonatkozó eredmények általában csak diszkrét idejű Markov-láncokra vannak megfogalmazva. A továbbiakban jelölje ξ_n a diszkrét, ξ_t pedig a folytonos idejű Markov-láncot.

A diszkrét idejű Markov-lánc olyan Markov-típusú sztochasztikus folyamat, amelynek indexparamétere $T = (0, 1, 2, \ldots)$. Azt mondjuk, hogy a folyamat az i állapotban van, ha $\xi_n = i$, ahol $n \in T$ és $i \in G$. Annak a valószínűségét, hogy ξ_{n+1} a j állapotba megy át, feltéve, hogy ξ_n az i állapotban van, egylépéses átmenetvalószínűségnek nevezzük és p_{ij}-vel jelöljük:

$$p_{ij} = \mathrm{P}(\xi_{n+1} = j | \xi_n = i).$$

A p_{ij} számokat mátrix formájában szokás elrendezni

$$\mathbf{P} = \begin{bmatrix} p_{01} & p_{02} & p_{03} & \cdots \\ p_{10} & p_{11} & p_{12} & \cdots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}.$$

Diszkrét idejű Markov-láncoknál feltesszük, hogy az egyes átmenetek között mindig ugyanannyi idő telik el. A $\mathbf{P}$ mátrixot a folyamat *átmenetvalószínűség mátrix*ának nevezzük, ahol az átmenetvalószínűségek eleget tesznek a következő összefüggéseknek:

(a) $p_{ii} = 0$ és $p_{ij} = 0$, ha i és j csúcsok nem szomszédosak,

(b) $p_{ij} \geq 0$, ha i és j csúcsok szomszédosak és

(c) $\sum_{j=0}^{\infty} p_{ij} = 1$, minden $i = 0, 1, 2, \ldots$

*Megjegyzés.*A $p_{ii} = 0$ feltétel általában nem szükséges, de mivel a folytatásban a folytonos idejű bolyongás definiálásakor szerepet játszik, ezért kimondjuk.

A diszkrét és a folytonos idejű Markov-láncok közötti megfeleltetés a következő konstrukció alapján történik. Képzeljük el, hogy minden $k \in G$ csúcsnál adottak a $\lambda_k > 0$ paraméterű Poisson folyamatok és a p_{kj} átmenetvalószínűségek. Ezen paraméterek által meghatározott folyamat a következőképpen működik: ha a folytonos idejű Markov-lánc a t időpontban a k csúcsban van, azaz $\xi_t = k$, akkor egységnyi idő helyett λ_k paraméterű exponencionális várakozási időután p_{kj} valószínűséggel vándorol a szomszédos j pontba. Az így kapott folyamat a következőképpen viselkedik: egységnyi időhelyet a véletlentől függő ideig a k állapotban található, majd ennek az időszaknak a végén p_{kj} valószínűségekkel átmegy valamelyik szomszédos j csúcsba, ahol ismét a véletlentől függő ideig tartózkodik és így tovább.

3.1. Definíció. Legyen ξ_t a fenti módón definiált folytonos Markov-lánc a $\mathcal{G}$ gráfon és legyen $\varphi : G \mapsto \mathbb{R}$ korlátos függvény. Ekkor $P^t\varphi(k) = \mathbb{E}\Big[\varphi(\xi_t)|\xi_0 = k\Big]$ a folyamat feltételes várható érték operátora, és a folyamat generátora

$$\mathcal{L}\varphi(k) = \lim_{t \to 0} \frac{1}{t}\Big(P^t\varphi(k) - \varphi(k)\Big).$$

Könnyű ellenőrizni a következő állítást, aminek egy kicsit bonyolultabb változatát még ebben a szakaszban igazoljuk.

3.1. ÁLLÍTÁS. *A fent leírt folytonos Markov-lánc generátora*

$$\mathcal{L}\varphi(k) = \lambda_k \sum_{j \neq k} p_{kj}\Big(\varphi(j) - \varphi(k)\Big). \tag{2}$$

Az is látható hogy $\mathcal{L}\phi(k) \leq 0$, ha k a ϕ lokális maximumának helye, vagyis teljesül a nevezetes *maximum elv.*

Mielőtt még rátérnénk a következő fejezetre, ezt a folyamatot tovább módosítjuk. Tegyük fel, hogy adott egy λ paraméterű Poisson folyamat úgy, hogy minden $\lambda_k \leq \lambda$. Azt mondjuk, hogy λ paraméterű exponenciális idő után a bolyongó részecske $\frac{\lambda-\lambda_k}{\lambda}$ valószínűséggel továbbra is marad a k csúcsban, illetve $\frac{\lambda_k}{\lambda}$ valószínűséggel ugrik valamely szomszédos csúcsba. Az így kapott folyamatot fogjuk a továbbiakban *folytonos Markov-láncként* emlegetni.

3.1. TÉTEL. *A fenti folytonos Markov-lánc generátorátorát a (2) képlet adja meg.*

Bizonyítás. Tegyük fel, hogy $\xi_t = k$. Jelölje A_t a λ paraméterű Poisson-folyamatot. Ekkor a $(t, t+s)$, $s > 0$, intervallumon a következőesemények lehetségesek:

I. $P(A_{t+s} - A_t = 0) = e^{-\lambda s}$,

II. $P(A_{t+s} - A_t = 1) = \lambda s e^{-\lambda s}$, ezen belül

(1) $P(\xi_{t+s} = k | \xi_t = k, A_{t+s} - A_t = 1) = \frac{\lambda-\lambda_k}{\lambda} \lambda s e^{-\lambda s}$,

(2) $P(\xi_{t+s} \neq k | \xi_t = k, A_{t+s} - A_t = 1) = \frac{\lambda_k}{\lambda} \lambda s e^{-\lambda s}$ és

III. $P(A_{t+s} - A_t \geq 2) = O(s^2)$.

Ekkor

$$(L\varphi)(k) = \lim_{s \to 0} \frac{\mathbb{E}(\varphi(\xi_{t+s}) | \xi_t = k) - \varphi(k)}{s},$$

ahol

$$\mathbb{E}(\varphi(\xi_{t+s}) | \xi_t = k) = \left(e^{-\lambda s} + \lambda s e^{-\lambda s} \frac{\lambda - \lambda_k}{\lambda} \right) \varphi(k) + \\ + \frac{\lambda_k}{\lambda} \lambda s e^{-\lambda s} \sum_{j \neq k} p_{kj} \varphi(j) + O(s^2).$$

Ebből adódóan

$$\mathcal{L}\varphi(k) = \lim_{s \to 0} \frac{\left(e^{-\lambda s} + \lambda s \frac{\lambda - \lambda_k}{\lambda} e^{-\lambda s} - 1 \right) \varphi(k)}{s} + \lambda_k e^{-\lambda s} \sum_{j \neq k} p_{kj} \varphi(j).$$

Felhasználva az összefüggést, amely szerint

$$1 = e^{-\lambda s} + \lambda s e^{-\lambda s} + O(s^2),$$

valamint az adódó egyszerűsítések és a határátmenet elvégzése után

$$\mathcal{L}\varphi(k) = \lambda_k \sum_{j \neq k} p_{kj} \Big(\varphi(j) - \varphi(k) \Big)$$

adódik, amit bizonyítani kellett. □

Ez a gondolatmenet azt is igazolja hogy az

$$u(t,k) := \mathbb{E}\Big[\varphi(\xi_t)|\xi_0 = k\Big] = P^t\varphi(k)$$

feltételes várható érték a $\partial_t u(t,k) = \mathcal{L}u(t,k)$ *Kolmogorov-egyenlet* megoldása. A maximum elv segítségével az is következik hogy a Kolmogorov-egyenletek korlátos megoldása egyértelmű, lásd [4], és a feltételes várható értékek meghatározzák a folyamat átmeneti valószínűségeit, tehát az általunk definiált két folytonos Markov-folyamat azonos.

A következő részben ennek az egyenletnek egy összetettebb formájával foglalkozunk, melynek megoldását a vizsgált gráfon a Feynman–Kac-képlet teszi lehetővé.

4. A Feynman–Kac-képlet

A 3. fejezetben definiált folytonos Markov-lánc és a Feynman–Kac-megoldóképlet alapján a következő tétel igazolható.

4.1. TÉTEL. *Legyenek adottak a* $\varphi : G \mapsto \mathbb{R}$ *korlátos és* $\gamma : \mathbb{R} \mapsto \mathbb{R}$ *felülről korlátos függvények. Ekkor a*

$$\partial_t u(t,k) = \mathcal{L}u(t,k) + \gamma(k)u(t,k) \tag{3}$$

differenciálegyenletnek az $u(0,k) = \varphi(k)$ *kezdeti értékéhez tartozó megoldását a Feynman–Kac-képlet adja:*

$$u(t,k) := \mathbb{E}\Big(\varphi(\xi_t)e^{\int_0^t \gamma(\xi_\tau)d\tau}\Big),$$

ahol $\xi_0 = k$.

Megjegyzés. A Feynman–Kac-megoldóképlet sokkal bonyolultabb módon definiált Markov-láncokra is alkalmazható.

Bizonyítás. A bizonyítást először a $t = 0$ helyen végezzük el. Vezessük be az

$$X(t) = \varphi(\xi_t) \quad \text{és}$$
$$Y(t) = e^{\int_0^t \gamma(\xi_\tau)d\tau}$$

jelöléseket. Ekkor

$$u(t,k) = \mathbb{E}(X(t)Y(t)).$$

Innen

$$\begin{aligned}\partial_t u(0,k) = &\lim_{t\to 0} \frac{\mathbb{E}\Big(X(t)Y(t)\Big) - \mathbb{E}\Big(X(0)Y(0)\Big)}{t} \\ &+ X(0) \lim_{t\to 0} \frac{\mathbb{E}\Big(Y(t) - Y(0)\Big)}{t} \\ &+ Y(0) \lim_{t\to 0} \frac{\mathbb{E}\Big(X(t) - X(0)\Big)}{t} \\ &+ \lim_{t\to 0} \mathbb{E} \frac{\Big(X(t) - X(0)\Big)\Big(Y(t) - Y(0)\Big)}{t}.\end{aligned}$$

Tudjuk, hogy

$$\begin{aligned}X(0) &= \varphi(\xi_0) = \varphi(k), \\ Y(0) &= e^{\int_0^0 \gamma(\xi_\tau)d\tau} = 1,\end{aligned}$$

és a fenti egyenletben a harmadik határérték Y természete miatt éppen nullával egyenlő, valamint

$$\partial_t Y(0) = e^{\int_0^0 \gamma(\xi_\tau)d\tau}\gamma(\xi_0) = \gamma(k) \quad \text{és}$$

$$\lim_{t\to 0} \frac{\mathbb{E}\Big(X(t) - X(0)\Big)}{t} = \lim_{t\to 0} \frac{\mathbb{E}\Big(\varphi(\xi_t)\Big) - \mathbb{E}\Big(\varphi(k)\Big)}{t} = \mathcal{L}\varphi(k).$$

A fentieket összefoglalva, a $t = 0$ helyen a t szerinti differenciálás elvégzése után az eredmény valóban a kívánt

$$\partial_t u(0,k) = u(0,k)\gamma(k) + \mathcal{L}\varphi(k).$$

Továbbá

$$u(t+s,k) = \mathbb{E}\Big(\varphi(\xi_{t+s})\alpha\beta\Big),$$

ahol

$$\alpha := e^{\int_0^t \gamma(\xi_\tau)d\tau}$$

és

$$\beta := e^{\int_0^s \gamma(\xi_{t+\tau})d\tau}.$$

Mivel a $u(t+s,k)$ értékét ugyanazzal az eljárással kapjuk a $u(t,\cdot)$ függvényből, mint $u(t,k)$-t $u(0,\cdot) = \varphi$-ből, ezért a 3. fejezetben definiált folytonos Markov-lánc $\mathcal{F}_t = \sigma\{\xi_u : u \le t\}$ természetes filtrációjára való tekintettel adódik, hogy

$$u(t+s,k) := \mathbb{E}\Big[\mathbb{E}\Big(u(s,\varphi(\xi_{t+s}))\beta\Big)\,|\,\mathcal{F}_t\Big].$$

Ebből adódóan a $t > 0$ időpontban az időszerinti differenciálást ugyanúgy lehet elvégezni, mint a $t = 0$ helyen.

$$\partial_t u(t,k) = \lim_{s\to 0} \frac{u(t+s,k) - u(t,k)}{s}$$
$$= \gamma(k)u(t,k) + \lim_{s\to 0} \frac{P^{t+s}\varphi - P^t\varphi}{s}.$$

Amiből a határátmenet végrehajtásával

$$\partial_t u(t,k) = \gamma(k)u(t,k) + \mathcal{L}u(t,k).$$

Ezzel a tétel bizonyítását befejeztük. □

4.2. TÉTEL. *A (3) egyenletrendszernek pontosan egy korlátos megoldása létezik.*

Bizonyítás. A tétel bizonyítását [4] tárgyalja. □

5. Végtelen rendszerek stabilitása

A 2. fejezetben már igazoltuk, hogy a véges egyenletrendszerek exponenciálisan stabilak. Azonban azt is láthattuk, hogy a stabilizáló hatás a rendszer méretének a növekedésével gyengül, melynek következtében az ott alkalmazott módszerek nagy rendszerek esetében nem adnak választ a stabilitás kérdésére.

A továbbiakban a végtelen rendszerek egyik nagy osztályát képviselő lineáris rendszerek stabilitás vizsgálatával foglalkozunk. Általánosabb eredmények nemlineáris rendszerekre időfüggő Markov-láncok elméletének alkalmazásával érhetők el, jelenlegi dolgozatunk nem tárgyalja e témakört. Nemlineáris rendszerekre vonatkozó stabilitási tételeket várhatóan e cikk folytatásaként közlünk.

A vizsgált feladat a következő:

$$\partial_t x(t,k) = \sum_{j\neq k} \lambda_k p_{kj}\Big(x(t,j) - x(t,k)\Big) - \gamma(k)x(t,k), \tag{4}$$

ahol

$$\begin{cases} \gamma(k) = 0, & k \neq 0 \\ \gamma(k) > 0 & k = 0. \end{cases}$$

Megjegyzés. Ha a fenti lineáris egyenletrendszerben nem szerepel a $\gamma(k)$ függvény, akkor éppen a 3. fejezetben definiált folytonos idejű Markov-lánc *Kolmogorov-egyenletét* kapnánk. Így viszont a Markov-lánchoz rendelt Feynman–Kac-képlethez

jutunk, amely természetesen lehetővé teszi a rendszer megoldását és annak kiértékelését is a vizsgált gráfon.

5.1. KÖVETKEZMÉNY. *Legyen adott egy tetszőleges $\mathcal{G} = (G, E)$ összefüggő gráf, ahol $|G| = \infty$. Ekkor $x_0 = x(0,k)$, $k \in G$, korlátos kezdeti érték feltétele mellett, a (4) egyenletrendszer megoldása a Feynman–Kac-képlet alapján*

$$x(t,k) = \mathbb{E}\left(x(0,\xi_t(k))\exp\left\{-\int_0^t \gamma(\xi_\tau(k))d\tau\right\}\right).$$

5.1. LEMMA. *Ha a vizsgált gráfon a folytonos Markov-lánc rekurrens, akkor*

$$\int_0^{+\infty} \gamma(\xi_\tau(k))d\tau \longrightarrow +\infty.$$

Bizonyítás. A bizonyítást a második *Borel–Cantelli*-lemmára alapozzuk, mely szerint ha $\sum P(A_k)$ divergens és az A_k események teljesen függetlenek, akkor 1 valószínűséggel végtelen sok A_k következik be.

Tudjuk, hogy folytonos idejű rekurrens Markov-lánc esetében a rendszer biztosan végtelen sokszor visszatér minden olyan állapotba, amelyet egyszer már elfoglalt, valamint, hogy a visszatérések egymástól függetlenek. Valamint, ha egy adott összefüggő gráf egy tetszőleges pontjából indítjuk a Markov-láncot, szintén a Borel–Cantelli-lemmával igazolható, hogy 1 valószínűséggel előbb vagy utóbb eléri a 0 állapotot, ahová azt követően szintén 1 valószínűséggel végtelen sokszor visszatér.

Jelölje c_k azt az időtartamot, amelyet a folyamat a 0 állapotban tölt a k-dik visszatérés alkalmával. Legyen C_k az az esemény, hogy $c_k \geq c > 0$. Ekkor

$$\sum_{\tau_k \leq t} P(C_k) \geq \sum_{\tau_k \leq t} e^{-\lambda c_k} = +\infty, \quad \text{ha} \quad t \to +\infty.$$

Következésképpen C_k eseményekből végtelen sok következik be, ahonnan azt kapjuk, hogy

$$\int_0^{+\infty} \gamma(\xi_\tau(k))d\tau = \gamma(0)\int_0^{+\infty} d\tau = \gamma(0)\sum_k^{\infty} c_k = +\infty. \qquad \square$$

5.2. LEMMA. *Ha a vizsgált gráfon a folytonos Markov-lánc tranziens, akkor*

$$\int_0^{+\infty} \gamma(\xi_\tau(k))d\tau < +\infty.$$

Bizonyítás. Ismert, hogy tranziens Markov-folyamat esetében a folyamat csak véges sok alkalommal tér vissza a 0 állapotba. Így az előző lemma bizonyításának alapján most

$$\sum_{k}^{\infty} c_k < +\infty,$$

amiből adódik, hogy

$$\int_0^{+\infty} \gamma(\xi_\tau(k))d\tau = \gamma(0) \int_0^{+\infty} d\tau = \gamma(0) \sum_{k}^{\infty} c_k < +\infty. \qquad \square$$

A fenti lemmák közvetlen következménye az alábbi tétel:

5.1. TÉTEL. *Legyen adott egy tetszőleges $\mathcal{G} = (G, E)$ összefüggő gráf, ahol $|G| = \infty$. Ekkor $x_0 = x(0, k)$, $k \in G$, korlátos kezdeti érték feltétele mellett, a (4) egyenletrendszer megoldása rekurrens Markov lánc esetében globálisan stabil.*

Tranziens Markov-lánc esetében a megoldás stabilitása nagyban függ a kezdeti értékektől, ami a következő példákkal illusztrálható. Tegyük fel, hogy az adott összefüggő gráf véges halmazán kívül a függvény kezdeti értéke nulla. Ekkor tranziens esetben is világossá válik a rendszer stabilitása. Ezzel szemben ha a függvény kezdeti értéke a gráf véges halmazán nulla a fennmaradó csúcsokban pedig szigorúan pozitív, a rendszer instabil.

6. Néhány példa, rekurrens és tranziens Markov-láncok

Végezetül három érdekes Markov-lánc rekurrencia tulajdonságát ismertetjük.

6.1. n-dimeziós szimmetrikus bolyongás

A klaszikus n-dimenziós szimmetrikus bolyongás állapottere az n-dimenziós euklideszi tér egész koordinátájú pontjaiból álló rács: azaz a rendszer állapota egy egész számokból álló $k = (k_1, k_2, \ldots, k_n)$ szám n-es. Szimmetrikus bolyongás eseténél minden irányba történő elmozdulás ugyanakkora valószínűséggel következik be.

6.1. TÉTEL. *(Pólya) Az egy- és kétdimenziós szimmetrikus bolyongásban a bolyongó részecske 1 valószínűséggel előbb vagy utóbb (tehát végtelen sokszor is) visszatér a kezdeti helyzetbe. Három dimezióban ez a valószínűség kisebb egynél.*

A bizonyítás megtalálható pl. [10]-ben. A bolyongáshoz tartozó egyenletrendszer a következő

$$\partial_t x(t, k) = \sum_{j \neq k} \frac{1}{2n} \Big(x(t, j) - x(t, k) \Big) - \gamma(k) x(t, k).$$

6.2. Sinai-féle bolyongás

Legyen a $p = \{p(k)\}$, $k \in Z$, ahol $0 < p(k) < 1$, C-beli egyenletes eloszlású független valószínűségi változó sorozat. Ekkor azt mondjuk, hogy a számegyenesen bolyongó részecske $x(n)$ az n pontból $p(n)$ $[1 - p(n)]$ valószínűséggel lép a tőle jobbra [balra] eső szomszédos pontba.

6.2. TÉTEL. *$x(n)$ tranziens, ha $\mathbb{E}\log(1 - p(x))/p(x) \neq 0$. $x(n)$ rekurrens, ha létezik konstans $c > 0$, úgy, hogy $p(k), (1 - p(k)) > c$ és teljesül*

$$\mathbb{E}\log((1 - p(x))/p(x)) = 0.$$

A példa további tárgyalása megtalálható [8]-ban. A Sinai-féle bolyongáshoz rendelt egyenletrendszer a következő

$$\begin{aligned}\partial_t x(t,k) = \lambda_k \big(&p(k)(x(t,k+1) - x(t,k)) + \\ &+ (1 - p(k))(x(t,k-1) - x(t,k))\big) - \gamma(k)x(t,k).\end{aligned}$$

6.3. Markov-lánc a számegyenesen

Legyen $X_0 = 0, X_1, X_2, \ldots$ Markov-lánc úgy, hogy

$$\begin{aligned}P(X_{n+1} = i + 1 | X_n = i) &= 1 - P(X_{n+1} = i - 1 | X_n = i) \\ &= \begin{cases} 1, & i = 0 \\ \frac{1}{2} + p_i, & i = 1, 2, \ldots, \end{cases}\end{aligned}$$

ahol $0 \leq p_i \leq \frac{1}{2}$, $i = 1, 2, \ldots$. $\{X_n\}$ egy olyan részecske mozgását leíró sorozat, amely 0-ból indulva a nemnegatív egész számokon keresztül nagyobb valószínűséggel távolodik a 0-tól, mint közeledik ahhoz. A feladat akkor válik érdekessé, ha a $\{p_i, i = 1, 2, \ldots\}$ sorozat 0-hoz tart, vagyis ha a 0 pont taszító ereje egyre inkább gyengül, ahogyan a részecske távolodik tőle.

6.3. TÉTEL. *Legyen X_n a fenti átmenetvalószínűségekkel adott Markov-lánc. Ekkor elég nagy i-re teljesül valamelyik a következőket eset közül:*

1. *$p_i \leq \frac{1}{4i} + O\left(\frac{1}{i^{1+\delta}}\right)$ $\delta > 0$, akkor X_i rekurrens.*
2. *Létezik $\theta > 1$ úgy, hogy $p_i \geq \frac{\theta}{4i}$, akkor X_i tranziens.*

A példa további tárgyalása megtalálható [1]-ben. A Markov-lánchoz rendelt egyenletrendszer $k \neq 0$ esetén

$$\partial_t x(t,k) = \lambda_k \left((1/2 + p_i)(x(t,k+1) + x(t,k-1)) - (1 + 2p_i)x(t,k)\right) - \gamma(k)x(t,k)$$

és

$$\partial_t x(t,0) = \lambda_k \left((x(t,1) + x(t,-1)) - 2x(t,0)\right) - \gamma(0)x(t,0).$$

Hivatkozások

[1] Csáki Endre, Földes Antónia, Révész Pál: *Transient NN random walk on the line*, Preprint: http://front.math.ucdavis.edu/author/A.Foldes.

[2] John G. Kemeny, J. Laurie Snell, Anthony W. Knapp: *Denumerable Markov Chains*, Springer-Verlag, 1976.

[3] Fritz József: *Valószínűségszámítás fizikusoknak*, Preprint: http://www.math.bme.hu/ jofri/JOFRI/OKTAT/index.html, Budapest, 1998.

[4] Kai Lai Chung: *Markov Chains With Stationary Transition Probabilities*, Springer-Verlag Berlin Heidelberg New York, 1967.

[5] Ralf Korn, Elke Korn: *Option Pricing and Portfolio Optimization*, American Mathematical Society, 2001.

[6] Rényi Alfréd: *Valószínűségszámítás*, Tankönyvkiadó, Budapest, 1968.

[7] Samuel Karlin, Howard M. Taylor: *Sztochasztikus folyamatok*, Gondolat, Budapest, 1985.

[8] Sinai, Ya. G.: *The limit behavior of a one-dimensional random walk in a random environment*, Teor. Veroyatnost i primenen, 1982, no. **2**, 247–258.

[9] Stewart N. Ethier, Thomas G. Kurtz: *Markov Processes. Characterization and Convergence,* Wiley, New York, 1986.

[10] William Feller: *Bevezetés a valószínűségszámításba és alkalmazásaiba*, Műszaki könyvkiadó, Budapest, 1978.

(Beérkezett: 2008. április 28.)

LUKIC ANIKÓ
BMGE, Differenciálegyenletek Tanszék
1111 Budapest, Egry József u. 1.
lukity@math.bme.hu

STABILITY THEORY FOR INFINITE SYSTEMS OF DIFFERENTIAL EQUATIONS

Anikó Lukic

In this paper the stability of infinite linear system of differential equations is concerned. It is supposed there is given a Markov chain on a countable graph associated to the system, and the solution is given by the Feynman-Kac formula. The aim of this research is to show that in case the Markov chain is recurrent, than the linear system of differential equation defined on the same graph is stable. At the end of the paper the transient case is considered too.

Alkalmazott Matematikai Lapok **26** (2009), 329-350.

EMISSZIÓS DISZKRÉT TOMOGRÁFIAI MÓDSZEREK ALKALMAZÁSA FAKTORSTRUKTÚRÁKRA

NAGY ANTAL

Az utóbbi években egy új fajta diszkrét tomográfiai probléma kutatása kezdődött el [3], amit *emissziós diszkrét tomográfiának*, röviden *EDT*-nek nevezünk. Ebben a modellben a teljes tér valamilyen homogén abszorbens anyaggal van kitöltve és a rekonstruálandó függvény egy tárgyat reprezentál, aminek a pontjai (radioaktív) sugárzást bocsátanak ki a környező térbe. A tárgy egy pontjából kibocsátott kezdeti aktivitás egy része az abszorbens anyagban elnyelődik a pont és a detektor távolságától függően. A rekonstrukció kiindulására szolgáló vetületek tehát nem tisztán az emisszióra vonatkozó adatokat tartalmazzák, hanem az abszorpció hatását is.

A fő célunk az volt, hogy olyan struktúrák térfogatát becsüljük meg néhány (jelen esetben 4) vetületből, melyeknek intenzitása az időben változik. Mindegyik vetület adott idő alatt lett elkészítve, melyek az adott pillanatban felvett képek sorozatából állt. A struktúrák vetületei először faktoranalízissel lettek elkülönítve a teljes vetületi adatsorozatot felhasználva. Mivel a faktoranalízis a struktúrák vetületeit csak egy szorzó konstans [4] erejéig képes meghatározni, ezért a korábbi heurisztikus módszer mellett egy új módszert is bemutatunk, mely az abszorpciós vetületek konzisztencia feltételén [2, 8] alapszik. Ezek után a mindegyik struktúrát külön-külön rekonstruáltuk az adott szorzó konstansokkal módosított faktor vetületekből diszkrét tomográfiai módszerrel. Az így kapott térfogatokat az adott struktúrák esetén összehasonlítottuk a cikkben az adott szorzó konstansokat meghatározó módszerek esetén.

1. Bevezetés

Néhány alkalmazásban csak az f függvény vetületeit lehet mérni. Ez gyakran előfordul például a nukleáris medicinában, ahol a rekonstruálandó objektum a radioaktív eloszlás valamely szervben, a vetületek pedig gamma kamerás felvételek különböző irányokból. Ilyen esetben Single Photon Emission Computed Tomography (SPECT) képalkotó módszerrel gyűjtik be az adott objektum tomográfiás szeleteinek a rekonstrukciójához szükséges adatokat.

Jelölje $f(r,t)$ a rekonstruálandó objektum radioaktivitásának intenzitás függvényét. Tegyük fel, hogy a térben az elnyelődés állandó és az elnyelődési együttható

$\mu \geq 0$ konstans mindenhol. A térbeli félegyenesek felírhatók

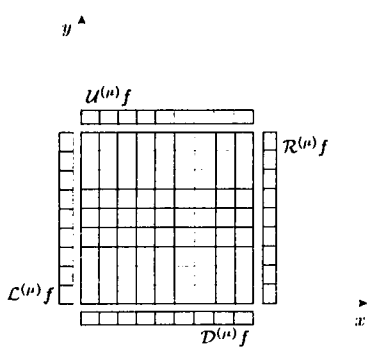
$$\ell(S, v) = \{S + u \cdot v \mid u \geq 0\}$$

alakban, ahol S a félegyenes kezdő pontja, illetve v az iránya. Így f *abszorpciós vetületét* $\ell(S, v)$ mentén a t időpillanatban a következőképpen lehet meghatározni

$$\left[\mathcal{P}^{(\mu)} f\right](S, v, t) = \int_0^\infty f(S + u \cdot v, t) \cdot e^{-\mu u}\, du\ .$$

Általában az abszorpciós vetületek értékeit párhuzamos félegyenesek mentén mérjük ugyanabban az időpillanatban (pl. vonal vagy sík detektorokat használva). Ilyen esetben v irányú (abszorpciós) vetületekről beszélünk. f rekonstrukcióját 4 *abszorpciós vetület* felhasználásával fogjuk végezni. A négy irány az egymással szemben lévő két-két vízszintes és függőleges irány. Megjegyezzük, hogy az emissziós tomográfiai modellben a szemközti vetületek általában nem határozhatók meg egymásból, ellentétben a klasszikus (transzmiszsziós) DT-ben használt modellel, ahol a szemközti vetületek tükörképei egymásnak.

Tegyük fel, hogy a négy vetületet a bal, jobb, felső és alsó irányokból vettük fel. Továbbá tegyük fel, hogy az f függvény értelmezési tartománya a 3-dimenziós egység kocka minden t időpillanatban (1. ábra).



1. ábra. A négy abszorpciós vetület elrendezése.

A *bal*, *jobb*, *felső* és az *alsó abszorpciós vetületeket* a következő formulákkal határozhatjuk meg

$$\left[\mathcal{L}^{(\mu)}f\right](y,z,t) = \left[\mathcal{P}^{(\mu)}f\right]((0,y,z),(1,0,0),t) = \int_0^1 f(u,y,z,t)\cdot e^{-\mu u}du\ , \tag{1.1}$$

$$\left[\mathcal{R}^{(\mu)}f\right](y,z,t) = \left[\mathcal{P}^{(\mu)}f\right]((1,y,z),(-1,0,0),t) = \int_0^1 f(1-u,y,z,t)\cdot e^{-\mu u}du\ , \tag{1.2}$$

$$\left[\mathcal{U}^{(\mu)}f\right](x,z,t) = \left[\mathcal{P}^{(\mu)}f\right]((x,1,z),(0,-1,0),t) = \int_0^1 f(x,1-u,z,t)\cdot e^{-\mu u}du\ , \tag{1.3}$$

$$\left[\mathcal{D}^{(\mu)}f\right](x,z,t) = \left[\mathcal{P}^{(\mu)}f\right]((x,0,z),(0,1,0),t) = \int_0^1 f(x,u,z,t)\cdot e^{-\mu u}du\ . \tag{1.4}$$

Másképpen megfogalmazva, az (1.1)–(1.4) egyenletek azt fejezik ki, hogy a detektorok az egységkocka bal, jobb, felső és alsó lapján fekszenek, a kocka felé néznek és az abszorpciós vetületeket olyan félegyenesek mentén mérik, melyek merőlegesek a kocka megfelelő oldalaira.

Az $f(r,t)$ függvény rekonstruálását három részben hajtjuk végre. Először szét kell bontani a 3D-s dinamikus objektumok eredeti vetületeit faktorstruktúrák vetületeire, majd a faktorstruktúrákhoz tartozó intenzitás értékeket határozzuk meg. Ezek után mindegyik faktorstruktúra 2D-s szeleteit rekonstruáljuk 4 vetületéből. Ezt a rekonstrukciót ismételve mindegyik szeletre megkapjuk a 3D-s rekonstruált faktorstruktúrákat.

2. Faktorstruktúrák

Tekintsük a következő problémát. Tegyük fel, hogy van egy 3D-s dinamikus tárgy, amelyet egy nemnegatív $f(r,t)$ függvénnyel ábrázolhatunk, ahol r és t jelöli rendre a térbeli pozíciót és az időt. Tegyük fel, hogy f felírható függvények lineáris kombinációjaként a következőképpen

$$f(r,t) = c_1(t)\cdot f_1(r) + c_2(t)\cdot f_2(r) + \cdots + c_K(t)\cdot f_K(r) + \eta(r,t), \tag{2.1}$$

ahol $k = 1,2,\ldots,K$, $(K \geq 1)$, $f_k(r)$ időben állandó 0, 1 értékű függvény, $c_k(t)$

a k-ik súly együttható, amely csak az időtől függ és $\eta(r,t)$ reprezentálja a zajt. Ismert, hogy f és η, továbbá f_i és f_j minden $i \neq j$-re korrelálatlanok.

2.1. A fantom

Az eljárást 3D fantom kísérlettel próbáltuk ki. A mi fantomunk — azaz az f függvény a (2.1) egyenletben — a vizelet kiválasztás egyszerűsített 3D-s matematikai modellje volt, amit Dr. Werner Backfrieder, AKH Vienna, Ausztria [1] biztosított számunkra. A modell 5 faktorból állt (azaz, $K = 5$), melyeket $f_1, f_2, \ldots, f_5$ jelölt a (2.1) egyenletben. Ezek a faktorok a két vérkeringési struktúrát, a két vese struktúrát és a húgyhólyagot ábrázolják. Mindegyik faktorstruktúra homogén, azaz az adott struktúrákat alkotó voxelek értéke minden időpillanatban 1, a struktúrák geometriai objektumokkal vannak megadva (diszkrét gömbökkel, hengerekkel, stb.). A faktorstruktúrák a 64^3 voxelből álló digitális térben vannak elhelyezve (a voxel mérete $6 \times 6 \times 6\,\mathrm{mm}$). A háttér a 64^3 voxel kocka maradéka. Az 1. táblázatban lévő adatok adnak további információt az adott struktúrákról.

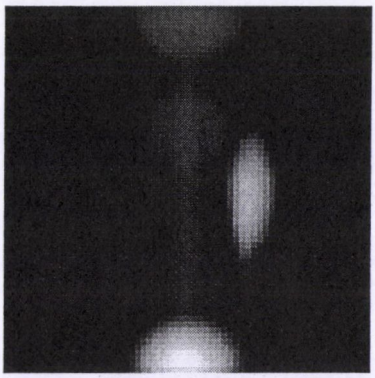
1. táblázat. A fantom struktúrái

Struktúra neve	térfogata (voxelben)
Szív és aorta	2652
Máj és lép	10603
Két vesekéreg	1350
Két vesemedence	606
Húgyhólyag	2094
Háttér	64^3 kocka maradéka

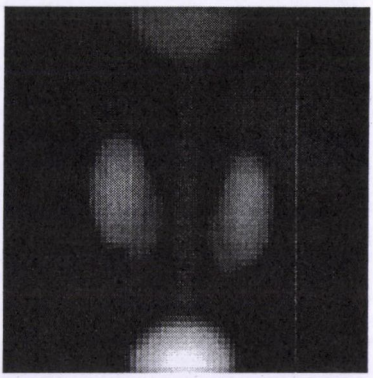
Annak érdekében, hogy a vetületi képek szimulálása egy nukleáris medicinai SPECT vizsgálat körülményeit kellő mértékben közelítse, abszorpciót, szórást, mélység függő felbontást, rész-térfogat hatást (partial volume effect) és még Poisson-zajt vettek figyelembe. A modellben $c_k(t)$ $(k = 1, 2, \ldots, 5)$ a faktor súlyok (azaz az intenzitások) az adott szervek működésének megfelelően időben változnak. A négy 64×64-es méretű abszorpciós vetületet ($\mathcal{L}^{(\mu)}f$, $\mathcal{R}^{(\mu)}f$, $\mathcal{U}^{(\mu)}f$ és $\mathcal{D}^{(\mu)}f$) 120 diszkrét időpillanatban állították elő. A 120 időpillanatban készült vetületből irányonkénti számított összegképek a 2. ábrán láthatóak.

2.2. Faktoranalízis

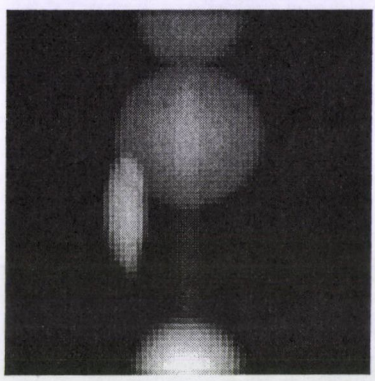
A 3D-s objektum mindegyik szimulált faktorstruktúrájának speciális dinamikája van (a radioaktivitás az idővel változik) a (2.1) egyenletnek megfelelően. Így egyes struktúrák vetületei faktoranalízissel elkülöníthetők a többi struktúrától. A faktoranalízist az [5, 6] publikációk szerint hajtották végre (Dr. Martin Samal,



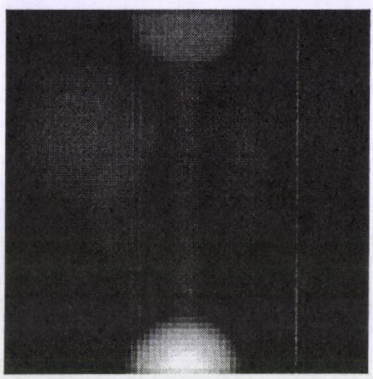
(a) A bal vetület összegképe



(b) A felső vetület összegképe



(c) A jobb vetület összegképe



(d) Az alsó vetület összegképe

2. ábra. 120 időpontban készült vetületek összegképei.

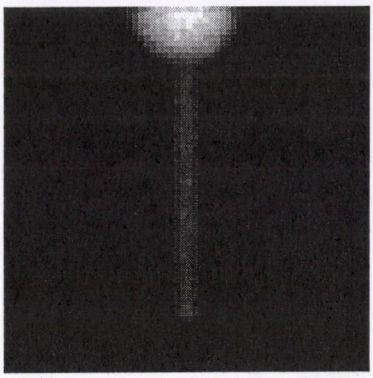
Charles University Prague, Csehország) mindegyik projekció sorozaton. A faktoranalízis eredménye 20 darab (azaz 4×5) 64×64 mátrix (a vetületi faktorok L_k, R_k, U_k és D_k betűkkel jelöltük) és a megfelelő súlyok

$$\left(c_k^{(l)}(t),\ c_k^{(r)}(t),\ c_k^{(u)}(t),\ \text{és}\ c_k^{(d)}(t)\right),\quad k = 1, 2, \ldots, 5.$$

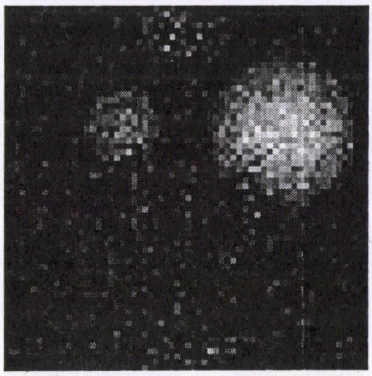
Példaként a „felső" irányból készült 5 képet mutatjuk be a 3. ábrán. A súlyok időbeli változását mutató 20 görbe

$$\left(c_k^{(l)}(t), c_k^{(r)}(t), c_k^{(u)}(t)\ \text{és}\ c_k^{(d)}(t),\ \text{ahol}\ k = 1, 2, \ldots, 5\right)$$

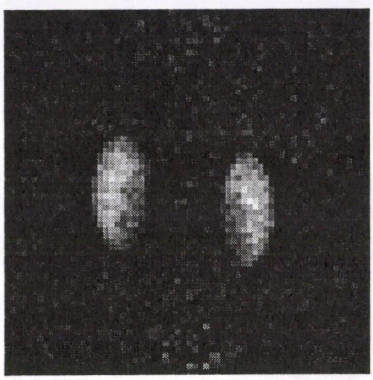
az 4. ábrán tekinthető meg.



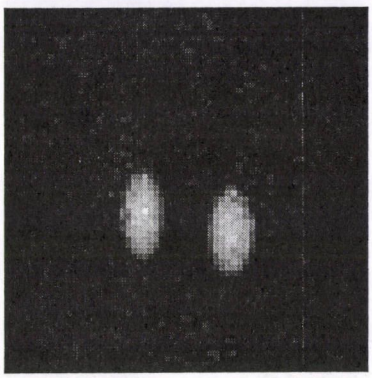
(a) Szív és aorta



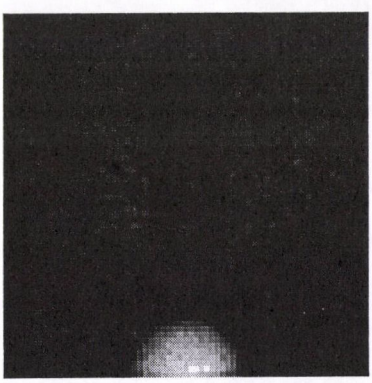
(b) Máj és lép



(c) Vesekéreg

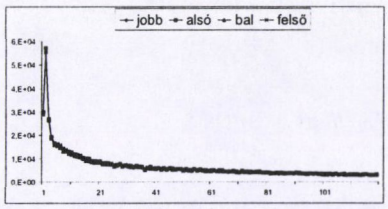


(d) Vesemedence

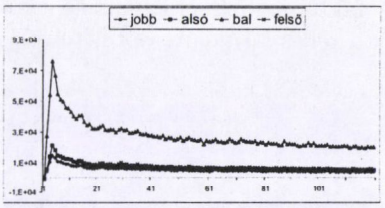


(e) Húgyhólyag

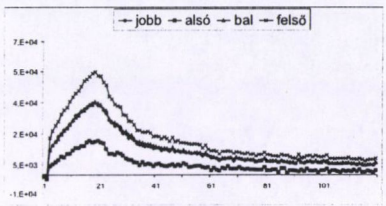
3. ábra. Az U_k képek, ahol $k = 1, 2, \ldots, 5$, ahogyan a „felső" $\mathcal{U}^{(\mu)} f$ vetületekből állt elő faktoranalízissel.



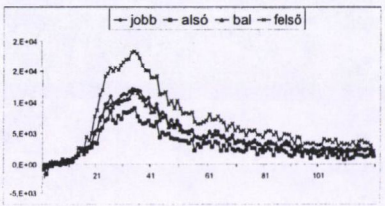
(a) Szív és aorta



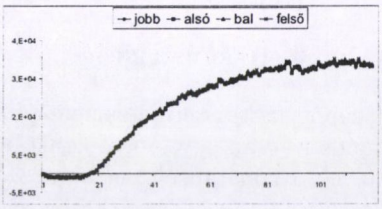
(b) Máj és lép



(c) Vesekéreg



(d) Vesemedence



(e) Húgyhólyag

4. ábra. A faktoranalízissel kapott súlyok görbéi.

Megjegyezzük, hogy a korrekt eljárás az lett volna, ha az azonos időpontban készült 4 vetületi képet egy képnek tekintették volna, és úgy hajtották volna végre a faktoranalízist. Eredményként olyan faktorokat kaptak volna, amelyek mindegyike 4 vetületi képet tartalmaz (tehát így is $4 \times 5 = 20$ vetületi képet), de csak 5 görbét $(c_1, \ldots, c_5)$! Ha ugyanis a (2.1) egyenletre alkalmazzuk a vetítő operátort, f_k vetületeiként megkapjuk az L_k, R_k, U_k és D_k vetületi képeket, amelyekhez így csak egyetlen $c_k(t)$ súly tartozik.

A mi esetünkben azonban minden egyes vetületre külön-külön végezték el a faktoranalízist. A faktoranalízis csak egy konstans szorzó erejéig tudja meghatározni $c_k(t)$-t és f_k-t, azaz az eredményként kapott $c_k(t)$ görbe és f_k faktor helyett más, $d \cdot c_k(t)$ görbe és f_k/d faktor állhat elő a faktoranalízis során $(d \neq 0)$. A faktoranalízis ily módon való végrehajtása azt eredményezte, hogy a különböző szorzók miatt pl. a májhoz és léphez tartozó (4(b) ábra) görbék eltérnek.

A faktoranalízissel kapott képek és az együtthatók a következő összefüggésben állnak a teljes struktúra vetületeivel:

$$\left[\mathcal{L}^{(\mu)}f\right](y,z,t) = \sum_{k=1}^{5} c_k^{(l)}(t)\cdot L_k(y,z) + \eta_L(y,z,t)\ ,$$

$$\left[\mathcal{R}^{(\mu)}f\right](y,z,t) = \sum_{k=1}^{5} c_k^{(r)}(t)\cdot R_k(y,z) + \eta_R(y,z,t)\ ,$$

$$\left[\mathcal{U}^{(\mu)}f\right](x,z,t) = \sum_{k=1}^{5} c_k^{(u)}(t)\cdot U_k(x,z) + \eta_U(x,z,t)\ ,$$

$$\left[\mathcal{D}^{(\mu)}f\right](x,z,t) = \sum_{k=1}^{5} c_k^{(d)}(t)\cdot D_k(x,z) + \eta_D(x,z,t)\ ,$$

ahol η_L, η_R, η_U és η_D jelöli a megfelelő maradékokat.

3. Rekonstrukció

A faktoranalízis során kapott faktor vetületek nem a bináris struktúrák abszorpciós vetületei. Azokat a módszer csak egy szorzó konstans erejéig képes meghatározni, ezért szükség van az adott konstansok meghatározására.

A faktorstruktúrák intenzitás értékeinek meghatározására (lásd 4. fejezet) adott heurisztikus módszernél a 3D faktor struktúrák reprezentatív szeleteit rekonstruáljuk különböző szorzó konstansokkal. Majd a legjobb rekonstrukciós eredményt adó szorzó konstanssal korrigált szeletek vetületeit rekonstruáljuk. A következőkben ismertetünk egy lehetséges módszert a abszorpciós vetületekből történő bináris mátrixok rekonstrukciójára.

3.1. Bináris mátrixok rekonstrukciója abszorpciós vetületeikből

Tekintsük az f_k szeletét $z = z_0$ magasságban. Az $f_k(x,y,z_0)$ szeletet egy bináris mátrixszal lehet ábrázolni, vagy ezzel ekvivalens módon egy $\xi = (\xi_1,\ldots,\xi_J) \in \{0,1\}^J$ vektorral, ahol ξ_j jelöli a j-edik elemét a mátrixnak, mondjuk sorfolytonos bejárásban, ahol $j = 0,1,\ldots,J$ és $J = n^2$.

Ismerve mindegyik f_k négy abszorpciós vetületét, f_k egy emissziós diszkrét tomográfiai eljárással (EDT) rekonstruálható [3]. Az EDT rekonstrukciós probléma egy lineáris egyenletrendszerrel írható le:

$$\mathbf{A}\xi = \mathbf{b}\ , \tag{3.1}$$

ahol $\mathbf{b} = (b_i)$, $i = 1,2,\ldots,I$ és $\mathbf{A}$ jelöli azt a mátrixot, amely ξ és $\mathbf{b}$ között adja meg az összefüggést. $\mathbf{A}$ elemei a vetületek geometriájából, illetve az ismert μ

abszorpciós együtthatóból kiszámíthatók. A $\mathbf{b}$ vektort méréssel kapjuk. Egy olyan bináris ξ vektort keresünk, amely kielégíti a (3.1) egyenletrendszert.

A zaj, a mérési hibák és a modell egyszerűsítése miatt nem remélhettük, hogy megtaláljuk a (3.1) egyenletet pontosan kielégítő ξ-t. A (3.1) egyenletet ezért célszerű egy optimalizálási problémaként átfogalmazni. Formálisan a következő célfüggvény minimumát kell megtalálni

$$C(\xi) = \|\mathbf{A}\xi - \mathbf{b}\| + \Psi_{\mathrm{sm}}(\xi), \tag{3.2}$$

ahol ξ bináris és

$$\Psi_{\mathrm{sm}}(\mathbf{x}) = \gamma_{\mathrm{sm}} \cdot \Phi_{\mathrm{sm}}(\mathbf{x}). \tag{3.3}$$

$\Phi_{\mathrm{sm}}(\mathbf{x})$-t következőképpen adjuk meg:

$$\Phi_{\mathrm{sm}}(\mathbf{x}) = \sum_{j=0}^{J-1} \sum_{\ell \in Q_j^m} g_{\ell,j} \cdot |x_j - x_\ell|, \tag{3.4}$$

ahol Q_j^m a j-edik pixel $m \times m$-es környezetében lévő pixelek indexeinek a halmaza, és $g_{\ell,j}$ a j-edik pixel köré rajzolt Gauss eloszlás (harang felület) ℓ-edik pixelben felvett értéke. A $g_{\ell,j}$ skalár az ℓ-edik és j-edik pixel távolságát súlyozza. A $\Psi_{\mathrm{sm}}(\mathbf{x})$ regularizációs kifejezést használva az optimalizálási algoritmust arra kényszerítjük, hogy olyan bináris mátrixot kapjunk megoldásul, amely az adott prototípus függvény alatt, lehetőleg nagy összefüggő homogén (csak 0-kból vagy csak 1-esekből álló) területeket tartalmaz.

A (3.2) egyenlet megoldásához a *homogén szimulált hűtés* optimalizálási módszert használtuk.

3.1.1. Homogén szimulált hűtés

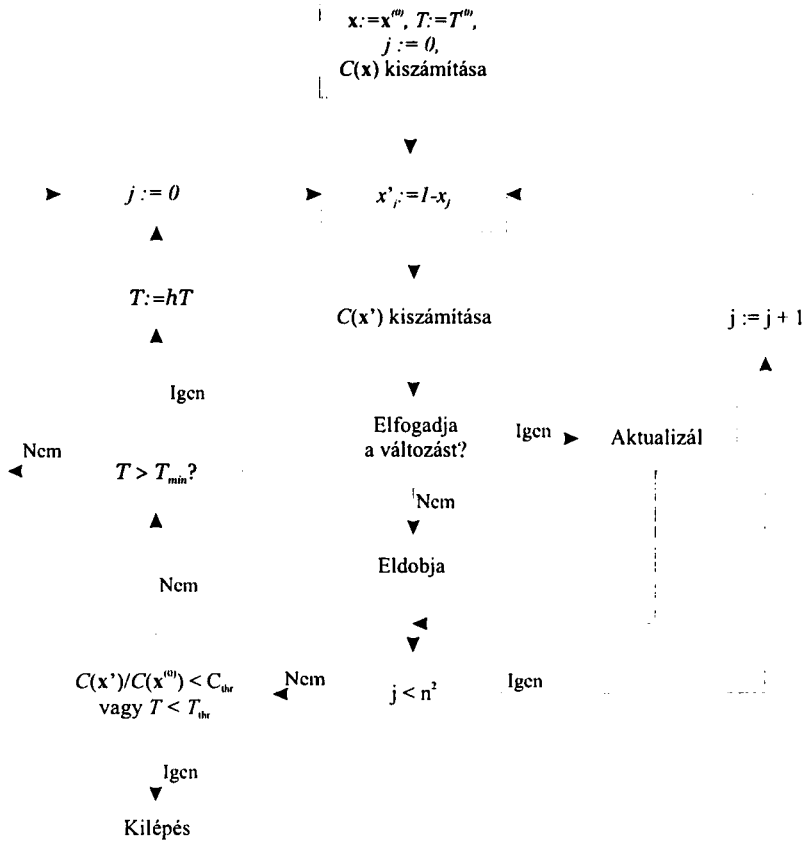
Az algoritmus egy tetszőleges $\mathbf{x}^{(0)}$ iniciális bináris képből és egy $T = T^{(0)}$ hőmérsékletről indul ki (5. ábra). Kiszámítja a $C(\mathbf{x})$ célfüggvény kezdő értékét. Egy iterációs lépésben az $\mathbf{x}$ vektor összes elemén végighaladva változtatja meg az adott elemet 0-ról 1-re, illetve 1-ről 0-ra.

Legyen $\mathbf{x}'$ az a kép, amely csak a j-edik pozícióban tér el az $\mathbf{x}$ képtől, azaz $x'_j = 1 - x_j$. Ezt a változtatást az algoritmus elfogadja, azaz $\mathbf{x}'$ lesz az új $\mathbf{x}$ vektor, ha $C(\mathbf{x}') < C(\mathbf{x})$. Ellenkező esetben a változtatást csak a $\Delta C = C(\mathbf{x}') - C(\mathbf{x})$ értékétől függő valószínűséggel fogadja el. Pontosan csak akkor, ha

$$\exp(-\Delta C/\kappa T) > z,$$

ahol κ a Boltzmann-állandó ($11.3805 \times 10^{-23}\ m^2 kgs^{-2}K^{-1}$), T az aktuális hőmérséklet, z pedig a $[0,1]$ intervallumon egyenletes eloszlású pszeudó véletlenszám-generátorral előállított szám.

Az iterációs lépések végén csökkentjük az aktuális hőmérsékletet és egy új iterációs lépést kezdünk. Az algoritmus akkor fejeződik be, ha az aktuális célfüggvény



5. ábra. A megvalósított homogén SA algoritmus folyamatábrája.

értékének és a célfüggvény kezdőértékének aránya egy előre megadott küszöbérték alá esik $\left(C(\mathbf{x}')/C\left(\mathbf{x}^{(0)}\right) < C_{\mathrm{thr}}\right)$ vagy az aktuális hőmérséklet egy adott hőmérsékleti pont alá kerül $(T < T_{\mathrm{thr}})$.

4. A faktorstruktúrák intenzitás értékeinek meghatározása

A vetületi mátrixokat nem tekinthetjük a bináris struktúrák abszorpciós vetületeinek, mert semmi sem biztosítja, hogy a faktorstruktúrákhoz tartozó voxelek

sugárzása egységnyi intenzitású. Ezért

$$c_k(t) \cdot \left[\mathcal{L}^{(\mu)} f_k\right](y,z) = c_k^{(l)}(t) \cdot L_k(y,z) ,$$
$$c_k(t) \cdot \left[\mathcal{R}^{(\mu)} f_k\right](y,z) = c_k^{(r)}(t) \cdot R_k(y,z) ,$$
$$c_k(t) \cdot \left[\mathcal{U}^{(\mu)} f_k\right](x,z) = c_k^{(u)}(t) \cdot U_k(x,z) ,$$
$$c_k(t) \cdot \left[\mathcal{D}^{(\mu)} f_k\right](x,z) = c_k^{(d)}(t) \cdot D_k(x,z) ,$$

minden $k = 1, 2, \ldots 5$-re. Ez azt jelenti, hogy a faktoranalízis a faktorok vetületeit csak egy szorzó konstans erejéig tudja meghatározni. Mielőtt bármilyen rekonstrukciós módszert használnánk, meg kell határoznunk a faktorstruktúrák valódi intenzitásait minden $k = 1, 2, \ldots, 5$-re.

$$d_k^{(l)} = c_k^{(l)}(t)/c_k(t), \qquad d_k^{(r)} = c_k^{(r)}(t)/c_k(t),$$
$$d_k^{(u)} = c_k^{(u)}(t)/c_k(t) \quad \text{és} \quad d_k^{(d)} = c_k^{(d)}(t)/c_k(t)$$

konstansokat minden $k = 1, 2, \ldots, 5$-re.

Két módszert adunk a faktorok intenzitásának meghatározására, mindegyik faktorra ugyanazt az eljárást használva.

4.1. Heurisztikus módszer

A módszer lényege [4] azon a megfigyelésen alapszik, hogy a célfüggvény értéke az adott rekonstrukciós eljárás során annál közelebb kerül a nullához, minél jobban megközelítjük a szorzó konstans értékét az adott faktor struktúra esetén.

4.1. Algoritmus. A faktorok intenzitás értékének meghatározása abszorpciós vetületekből

Bemenet: A faktor abszorpciós vetületei
Kimenet: A faktor intenzitás értéke

1. lépés: Válasszuk ki az adott faktor abszorpciós vetületi kép sorozatából azt a reprezentatív szeletet, amelynek a legnagyobb az összértéke.

2. lépés: Rekonstruáljuk a 3D-s faktorstruktúrának a reprezentatív szeletét különböző λ szorzót használva a (4.1) célfüggvény minimalizálásánál.

$$C(\lambda) = \|\mathbf{A} \cdot (\lambda \cdot \xi) - \mathbf{b}\| + \Psi_{\mathrm{sm}}(\xi), \tag{4.1}$$

ahol ξ a bináris faktor reprezentatív szeletét leíró vektor és a $\Psi_{\mathrm{sm}}(\xi)$ a 3.1. fejezetben bevezetett regularizációs kifejezés.

3. lépés: A különböző kipróbált λ szorzók közül a legkisebb $C(\lambda)$ célfüggvényhez tartozót választottuk ki λ értékének, azaz $\lambda^* = \arg\min_{\lambda}\{C(\lambda)\}$.

VÉGE

4.2. Konzisztencia feltételen alapuló módszer

A szorzó konstans egy másik lehetséges meghatározása, ha az abszorpciós vetületekre vonatkozó konzisztencia feltételt [2, 8] használjuk.

Az algoritmus ismertetéséhez a következő definíciókra van szükségünk. A $\mathcal{P}_X^{(\mu)}$, $\mathcal{P}_Y^{(\mu)}$, $\mathcal{P}_X$ és $\mathcal{P}_Y$ abszorpciós, illetve abszorpció mentes vetületeket definiáljuk a következőképpen a Q mérhető síkhalmazra:

$$\left[\mathcal{P}_X^{(\mu)}Q\right](y) = \int_{-\infty}^{\infty} \chi_Q(x,y)e^{-\mu x}dx,$$

$$\left[\mathcal{P}_Y^{(\mu)}Q\right](x) = \int_{-\infty}^{\infty} \chi_Q(x,y)e^{-\mu y}dy,$$

$$[\mathcal{P}_X Q](y) = \int_{-\infty}^{\infty} \chi_Q(x,y)dx,$$

$$[\mathcal{P}_Y Q](x) = \int_{-\infty}^{\infty} \chi_Q(x,y)dy,$$

ahol χ_Q jelöli a $Q \in \mathbb{R}^2$ karakterisztikus függvényét.

Jelöljük L, U, R, D-vel az adott faktorstruktúrák vetületeit. Ezek után definiáljuk az i-edik keresztmetszeten a második és harmadik vetületeket a [8]-nak megfelelően a következőképpen:

$$f^i_{LU}(x) = \mathcal{P}_Y^{(\mu)}\{(x,y)|L(1-y,z=i) \geq x\},$$
$$f^i_{UL}(y) = \mathcal{P}_X^{(\mu)}\{(x,y)|U(x,z=i) \geq y\},$$
$$f^i_{UR}(x) = \mathcal{P}_X^{(\mu)}\{(x,y)|U(1-x,z=i) \geq y\},$$
$$f^i_{RU}(y) = \mathcal{P}_Y^{(\mu)}\{(x,y)|R(1-y,z=i) \geq x\},$$
$$f^i_{RD}(x) = \mathcal{P}_Y^{(\mu)}\{(x,y)|R(y,z=i) \geq x\},$$
$$f^i_{DR}(y) = \mathcal{P}_X^{(\mu)}\{(x,y)|D(1-x,z=i) \geq y\},$$
$$f^i_{DL}(x) = \mathcal{P}_X^{(\mu)}\{(x,y)|D(x,z=i) \geq y\},$$
$$f^i_{LD}(y) = \mathcal{P}_Y^{(\mu)}\{(x,y)|L(y,z=i) \geq x\},$$

és

$$f^i_{ULY}(x) = \mathcal{P}_Y\{(x,y)|f^i_{UL}(y) \geq x\},$$
$$f^i_{RUY}(x) = \mathcal{P}_Y\{(x,y)|f^i_{RU}(y) \geq x\},$$
$$f^i_{DRY}(x) = \mathcal{P}_Y\{(x,y)|f^i_{DR}(y) \geq x\},$$
$$f^i_{LDY}(x) = \mathcal{P}_Y\{(x,y)|f^i_{LD}(y) \geq x\},$$

ahol az L, U, R, D indexek sorozatai azt jelzik, hogy milyen sorrendben végeztük el a vetületek további vetítését.

Végül, vegyük a rekonstruálandó faktor i-edik szeletének második, illetve harmadik vetületének integrálját mind a négy irányból, és jelöljük azokat a következőképpen:

$$F^i_{LU}(c) = \int_0^c f^i_{LU}(x)dx, \qquad F^i_{UL}(c) = \int_0^c f^i_{ULY}(x)dx\ ,$$

$$F^i_{UR}(c) = \int_0^c f^i_{UR}(x)dx, \qquad F^i_{RU}(c) = \int_0^c f^i_{RUY}(x)dx\ ,$$

$$F^i_{RD}(c) = \int_0^c f^i_{RD}(x)dx, \qquad F^i_{DR}(c) = \int_0^c f^i_{DRY}(x)dx\ ,$$

$$F^i_{DL}(c) = \int_0^c f^i_{DL}(x)dx, \qquad F^i_{LD}(c) = \int_0^c f^i_{LDY}(x)dx\ .$$

4.2. Algoritmus. A faktorok intenzitás értékének meghatározása abszorpciós vetületekből

Bemenet: A faktor abszorpciós vetületei
Kimenet: A faktor intenzitás értéke

1. lépés: Korrekció. Határozzuk meg azokat az α, β és γ értékeket, amelyekre a (4.2) kifejezés minimális minden $c > 0$ értékére.

$$\begin{aligned}
&\sum_i \left((\alpha \cdot F^i_{LU}(c) - F^i_{UL}(c))^2 \cdot \sqrt{F^i_{LU}(c) \cdot F^i_{UL}(c)} \right) \quad + \\
&\sum_i \left((\alpha \cdot F^i_{LD}(c) - \beta \cdot F^i_{DL}(c))^2 \cdot \sqrt{F^i_{LD}(c) \cdot F^i_{DL}(c)} \right) \quad + \\
&\sum_i \left((\beta \cdot F^i_{DR}(c) - \gamma \cdot F^i_{RD}(c))^2 \cdot \sqrt{F^i_{DR}(c) \cdot F^i_{RD}(c)} \right) \quad + \\
&\sum_i \left((\gamma \cdot F^i_{RU}(c) - F^i_{UR}(c))^2 \cdot \sqrt{F^i_{RU}(c) \cdot F^i_{UR}(c)} \right) \quad \to min.
\end{aligned} \tag{4.2}$$

A kapott α, β és γ értékekkel módosítsuk az F-eket minden i-re:

$$\hat{F}^i_{LU}(c) = \alpha \cdot F^i_{LU}(c), \qquad \hat{F}^i_{LD}(c) = \alpha \cdot F^i_{LD}(c),$$
$$\hat{F}^i_{DL}(c) = \beta \cdot F^i_{DL}(c), \qquad \hat{F}^i_{DR}(c) = \beta \cdot F^i_{DR}(c),$$
$$\hat{F}^i_{RD}(c) = \gamma \cdot F^i_{RD}(c), \qquad \hat{F}^i_{RU}(c) = \gamma \cdot F^i_{RU}(c).$$

2. lépés: Intenzitás érték meghatározása egy adott szeletre. Keressük meg azokat a maximális ρ^i_{LU}, ρ^i_{UR} , ρ^i_{RD} és ρ^i_{DL} értékeket, melyekre teljesül a

konzisztencia feltétel [8] minden $c > 0$ értékre, azaz

$$\begin{aligned}
&\text{ha } \hat{F}^i_{LU}(c) < \hat{F}^i_{UL}(c), \quad \text{akkor} \quad \hat{F}^i_{LU}\left(\rho^i_{LU} \cdot c\right) \geq \hat{F}^i_{UL}(c),\\
&\text{ha } \hat{F}^i_{UR}(c) < \hat{F}^i_{RU}(c), \quad \text{akkor} \quad \hat{F}^i_{UR}\left(\rho^i_{UR} \cdot c\right) \geq \hat{F}^i_{RU}(c),\\
&\text{ha } \hat{F}^i_{RD}(c) < \hat{F}^i_{DR}(c), \quad \text{akkor} \quad \hat{F}^i_{RD}\left(\rho^i_{RD} \cdot c\right) \geq \hat{F}^i_{DR}(c),\\
&\text{ha } \hat{F}^i_{DL}(c) < \hat{F}^i_{LD}(c), \quad \text{akkor} \quad \hat{F}^i_{DL}\left(\rho^i_{DL} \cdot c\right) \geq \hat{F}^i_{LD}(c).
\end{aligned}$$

A $\rho^i_{LU}, \rho^i_{UR}, \rho^i_{RD}, \rho^i_{DL}$ közül válasszuk ki a maximális értéket, jelölje az i-edik szeletre kapott értéket λ^i, azaz

$$\lambda^i = \max(\rho^i_{LU}, \rho^i_{UR}, \rho^i_{RD}, \rho^i_{DL}).$$

VÉGE

Az 1. lépéssel azt próbáljuk elérni, hogy a második vetületek integráljai lehetőleg közel azonosak legyenek a [8] 2.3-as tétel (24)-es egyenletének megfelelően.

A 2. lépés-ben a [8] 2.3-as tétel (25)-ös és (27)-es egyenleteinek megfelelően egy olyan szorzó konstanst határozunk meg az adott faktorstruktúrára, amely azt biztosítja, hogy létezik egy olyan mérhető síkhalmaz, melyeket az adott vetületek határoznak meg.

A 4.2. algoritmust természetesen csak olyan metszetekre érdemes használni, amelyekben az adott faktor jelen van. Az ilyen metszetek kiválasztására egy egyszerű küszöbölést választhatunk. Például csak azokra a szeletekre számoljuk ki a λ-t, amelyekre az alábbi összeg elér egy adott küszöböt:

$$\mathrm{SUM}^i = \int_{-\infty}^{\infty} L(y, z = i) + \int_{-\infty}^{\infty} R(y, z = i) + \int_{-\infty}^{\infty} U(x, z = i) + \int_{-\infty}^{\infty} D(x, z = i). \quad (4.3)$$

5. Eredmények

A két módszerrel meghatározott intenzitás értékeket a következő táblázatok tartalmazzák.

A rekonstrukció során a rekonstruálandó objektumok mérete miatt $m = 3$-at választottuk a (3.4) egyenletben szereplő Q^m_j szomszédság számára. A (3.3) egyenletben szereplő γ_{sm} regularizációs skalár esetében szintén figyelembe kellett vennünk a struktúrák méretét, illetve a torzítások hatását (pl. zaj) is. Az egyes struktúráknál, mind a két módszer esetén ugyanazokat a γ_{sm} értékeket használtuk a rekonstrukció során. A helyreállított reprezentatív szeletek átlag képei a 6., 7., 8., 9. és a 10. ábrákon láthatók.

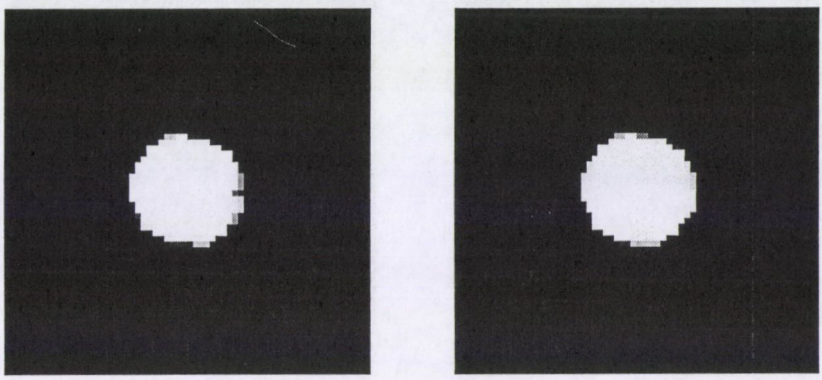
A rekonstrukciós eljárás megismétlésével a sztochasztikus módszer miatt más és más eredményt kaphatunk. A teljes rekonstrukciós eljárást 100-szor megismételtük mindegyik struktúrára azért, hogy információt kapjunk a megismételt rekonstrukciók különbségeiről. Az átlagos térfogatokat a megismételt rekonstrukciós

2. táblázat. A heurisztikus módszerrel kapott vetületi skalár értékek.

Struktúra neve	Szorzó konstans értéke
Szív és aorta	353.95
Máj és lép	23.65
Két vesekéreg	145.0
Két vesemedence	172.0
Húgyhólyag	206.5

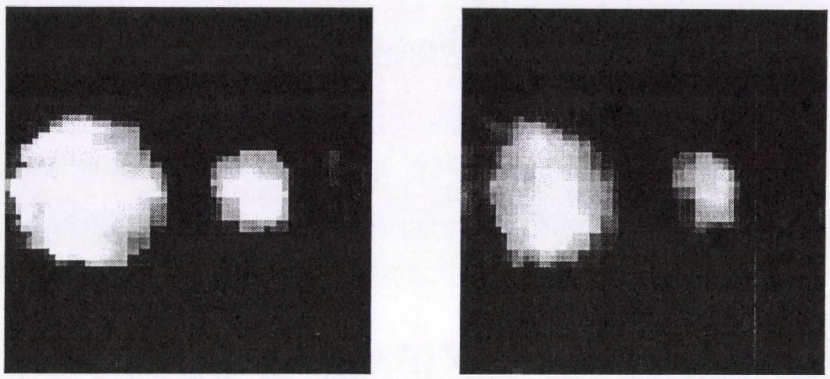
3. táblázat. A konzisztencia feltétellel kapott vetületi skalár értékek.

Struktúra neve	Szorzó konstans értéke
Szív és aorta	399.67
Máj és lép	32.03
Két vesekéreg	72.54
Két vesemedence	102.06
Húgyhólyag	195.7



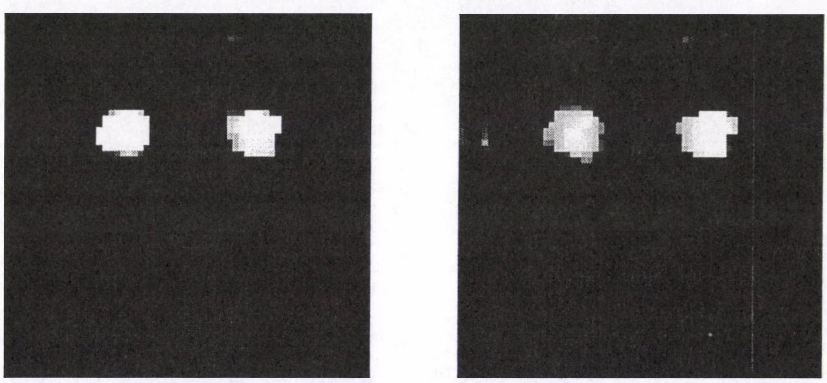
(a) Heurisztikus módszer (b) Konzisztencia módszer

6. ábra. A szív és aorta reprezentatív szeletének átlag képe a kétféle módszer alapján.



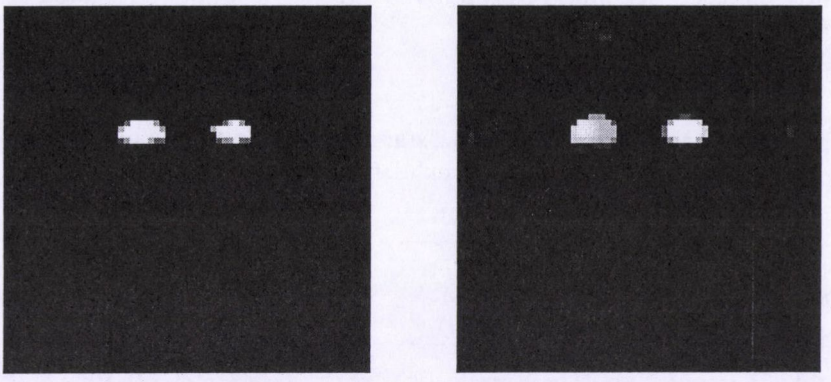
(a) Heurisztikus módszer (b) Konzisztencia módszer

7. ábra. A máj és a lép reprezentatív szeletének átlag képe a kétféle módszer alapján.



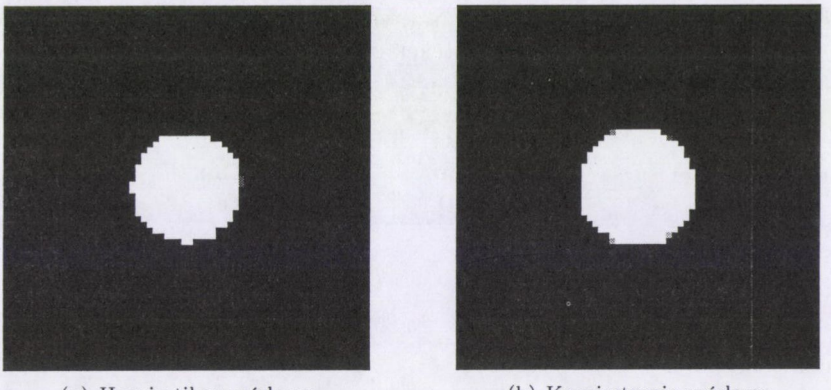
(a) Heurisztikus módszer (b) Konzisztencia módszer

8. ábra. A vesekérgek reprezentatív szeletének átlag képe a kétféle módszer alapján.



(a) Heurisztikus módszer (b) Konzisztencia módszer

9. ábra. A vesemedencék reprezentatív szeletének átlag képe a kétféle módszer alapján.



(a) Heurisztikus módszer (b) Konzisztencia módszer

10. ábra. A húgyhólyag reprezentatív szeletének átlag képe a kétféle módszer alapján.

eredményekből számítottuk. A 4. és 5. táblázatokban a rekonstruált térfogatoknak az eredeti térfogathoz viszonyított százalékos arányát is meghatároztuk (a zárójelben lévő számok). Az utolsó oszlop a 100-szor helyreállított térfogatok szórását mutatja.

4. táblázat. A rekonstruált struktúrák statisztikai eredményei a heurisztikus módszerrel meghatározott intenzitás értékek esetén.

Struktúra neve	Eredeti térfogat (voxel)	Helyreállított struktúra (voxel)	Szórás(voxel)
Szív és aorta	2652	2541 (96 %)	5.29
Máj és lép	10603	9486 (89 %)	100
Vesekérgek	1350	1450 (107 %)	17.1
Vesemedencék	606	511 (84 %)	5.4
Húgyhólyag	2094	1925 (92 %)	3.95

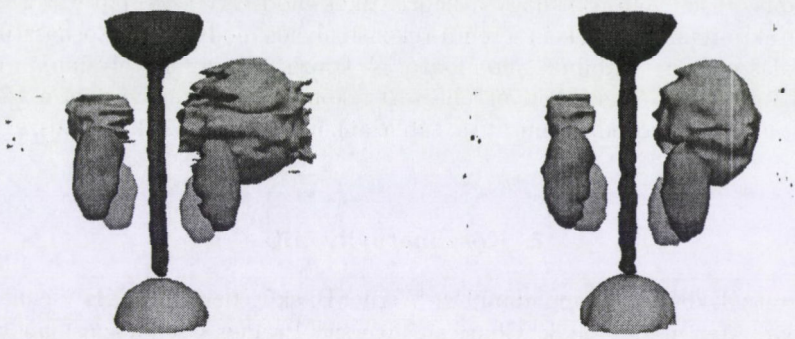
5. táblázat. A rekonstruált struktúrák statisztikai eredményei a konzisztencia feltétellel meghatározott intenzitás értékek esetén.

Struktúra neve	Eredeti térfogat (voxel)	Helyreállított struktúra (voxel)	Szórás(voxel)
Szív és aorta	2652	2657 (100 %)	13.88
Máj és lép	7023	9486 (66 %)	86
Vesekérgek	1350	1570 (116 %)	39.97
Vesemedencék	606	559 (92 %)	29.96
Húgyhólyag	2094	2267 (108 %)	26.63

A megismételt rekonstrukciókból kapott átlag szeleteket mindegyik struktúrára egy alkalmas vágási értéket használva jelenítettük meg a Slicer szoftver [7] használatával (11. és 12. ábrák).

(a) Heurisztikus módszer (b) Konzisztencia módszer

11. ábra. A rekonstruált átlagos struktúrák 3D-s megjelenítése elölnézetben a kétféle módszer alapján.



(a) Heurisztikus módszer (b) Konzisztencia módszer

12. ábra. A rekonstruált átlagos struktúrák 3D-s megjelenítése hátsó nézetben a kétféle módszer alapján.

6. Diszkusszió

Az EDT egy SPECT-beli lehetséges alkalmazását mutattuk be. Három lépéses eljárást javasoltunk a dinamikus vese SPECT vizsgálatból nyert 3D-s faktorstruktúrák 4 abszorpciós vetületből történő rekonstruálására. Az első lépésben a faktorstruktúrák faktoranalízissel lettek szétválasztva egymástól, majd a második lépésben két módszer (heurisztikus és konzisztencia feltételen alapuló) segítségével határoztuk meg a struktúrák intenzitás értékét. Végül, a harmadik lépésben egy EDT módszerrel rekonstruáltuk a 3D-s struktúrákat.

A faktorstruktúrák intenzitás értékének meghatározásakor a heurisztikus és a konzisztencia feltételen alapuló algoritmus esetén különböző szorzó konstansokat kaptunk eredményül. Ennek megfelelően a rekonstruált struktúrák is eltérnek.

Megfigyeltük, hogy a rekonstruált máj és lép struktúrák kevésbé voltak simák, mint a többi faktorstruktúrák. A hiba forrása az lehet, hogy az abszorpció miatt a máj aszimmetrikusan elhelyezkedő nagy részét bizonyos irányokból csak részlegesen lehet látni (például a bal oldali vetület a 2. ábrán). Hasonló magyarázat adható a vesemedencék és a húgyhólyag esetében is (4. táblázat).

A vetületek kiszámításakor az abszorpciót figyelembe vettük a rekonstrukcióban, de nem hajtottunk végre semmilyen szórás-, illetve zaj-korrekciót. A probléma egyszerűbb lenne, ha a faktoranalízist csak egy képsorozaton hajtanánk végre (a 4 vetület sorozatai helyett), ahol egy kép négy megfelelő vetület kompozíciója lenne.

Mivel az eredeti modell nem áll rendelkezésünkre, így csak az adott struktúrák térfogataihoz tudjuk hasonlítani az eredményeinket, amely csak részben ad pontos képet a rekonstrukció pontosságáról. Azt a következtetést vonhatjuk le, hogy a faktorstruktúrák helyreállított térfogatai nincsenek messze az igazi értékektől. A 4. táblázat azt mutatja, hogy a heurisztikus módszerrel meghatározott faktorstruktúrák intenzitás értékei esetén a rekonstrukciós módszer a sztochasztikus optimalizálás ellenére stabilnak mondható. A konzisztencia feltétel alapján meghatározott intenzitás értékekkel végrehajtott rekonstrukció után viszont a számított térfogat nagyobb szóródást mutat (5. táblázat), mint a heurisztikus módszer esetén.

7. Köszönetnyilvánítás

Szeretnék köszönetet mondani Dr. Werner Backfriedernek (AKH Vienna, Ausztria) és Dr. Martin Samalnak (Charles University Prague, Csehország), hogy a cikk megírásához adatokat bocsátottak rendelkezésemre. Külön szeretnék köszönetet mondani Kuba Attilának, aki halála előtt, időt és fáradtságot nem kímélve irányította, megjegyzéseivel és javaslataival segítette munkámat. Kutatásomat az OTKA T048476 pályázat is támogatta.

Hivatkozások

[1] Backfrieder, W., Samal, M., and Bergmann, H.: *Towards estimation of compartment volumes and radionuclide concentrations in dynamic SPECT using factor analysis and limited number of projections.* Physica Medica **15**, 3 (1999), 160.

[2] Kuba, A.: *Reconstruction of measurable sets from two absorbed projections.* Technical Report of Dept. of Computer Science, Univ. of Szeged (2005).

[3] Kuba, A., and Nivat, M.: *Reconstruction of discrete sets from absorbed projections.* In Proceedings of the 9th International Conference, Discrete Geometry for Computer Imagery (Berlin, 2000), G. Borgefors, I. Nyström, and G. Sanniti di Baja, Eds., vol. 1953 of Lecture Notes in Computer Sciences, Springer Verlag, pp. 137–148.

[4] Nagy, A., Kuba, A., and Samal, M.: *Reconstruction of factor structures using discrete tomography method.* Electronic Notes in Discrete Mathematics **20** (2005), 519–534.

[5] Samal, M., Karny, M., Surova, H., Marikova, E., and Dienstbier, Z.: *Rotation to simple structure in factor analysis of dynamic radionuclide studies.* Phys. Med. Biol. **32** (1987), 371–382.

[6] Samal, M., Nimmon, C. C., Britton, K. E., and Bergmann, H.: *Relative renal uptake and transit time measurements using functional factor images and fuzzy regions of interest.* Eur. J. Nucl. Med. **25**, 1 (1998), 48–54.

[7] http://www.slicer.org.

[8] Zopf, S., and Kuba, A.: *Reconstruction of measurable sets from two generalized projections.* Electronic Notes in Discrete Mathematics **20** (2005), 47–66.

(Beérkezett: 2008. április 30.)

NAGY ANTAL
Szegedi Tudományegyetem
Képfeldolgozás és Számítógépes Grafika Tanszék
6701 Szeged Pf. 652
nagya@inf.u-szeged.hu

APPLYING EMISSION DISCRETE TOMOGRAPHY METHODS ON FACTOR STRUCTURES

Antal Nagy

First, consider the following problem. Let us suppose that there is a 3D dynamic object, which can be represented by a non-negative function $f(r,t)$, where r and t denote the position in space and time, respectively. Suppose that f can be expressed as a weighted composite of a

set of (so far unknown) binary valued functions $f_k(r)$, $k = 1, 2, \ldots, K$ ($K \geq 1$) being constant in time, such that

$$f(r,t) = c_1(t) \cdot f_1(r) + c_2(t) \cdot f_2(r) + \cdots + c_K(t) \cdot f_K(r) + \eta(r,t), \tag{7.1}$$

where $c_k(t)$ denote the k-th weighting coefficient, which depends on time, and $\eta(r,t)$ represents the noise or residual in (r,t). Given the assumption that η and f are uncorrelated, $c_k(t)$ and $f(r)$ are to be determined such that f_i are independent from f_j for all $i \neq j$. If the values of $f(r,t)$ are available then the problem can be solved by factor analysis.

However, it can happen that we cannot measure the function f in the points of the space, but we can measure certain projections only. This is frequently the case, for example, in nuclear medicine, where the dynamic object is the radioactivity distribution in some human organ and the projections are gamma camera images from different directions. In this case SPECT imaging is applied to reconstruct the cross-sections of the object.

Let $f(r,t)$ denote the intensity function of the object to be reconstructed. Suppose that the absorption in the space is constant, that is the absorption coefficient is $\mu \geq 0$ everywhere. All half-lines in the space can be described as $\ell(S,v) = \{S + u \cdot v \mid u \geq 0\}$, where S and v are the point and direction of the half-line, respectively. Then the projections of f in time t can be measured along $\ell(S,v)$ half-lines by point detectors as follows

$$[\mathcal{P}^{(\mu)} f](S,v,t) \;=\; \int_0^\infty f(S + u \cdot v, t) \cdot e^{-\mu u}\, du\ .$$

Usually, the absorbed projection values are measured along many parallel half-lines simultaneously (e.g., by using line or plane detectors).

The method was tested on 3D phantom experiment. Our phantom (i.e., the function f in (7.1) was a simplified 3D mathematical model of the human renal system (it was provided by Dr. Werner Backfrieder, AKH Vienna, Austria). Each simulated factor structure of the whole 3D object had specific dynamics (radioactivity changes with time) according to (7.1), so, their projections seemed to be separable from the projections of other structures by factor analysis. The factor analysis was performed on each sequence of projections by the method published in [5, 6] using spatial constraints (Dr. Martin Samal, Charles University Prague, Czech Republic).

The projection images cannot be considered as the absorbed projections of the factor. However, the absorbed projections of the factor structures can be computed from these images by suitable multiplications. Therefore, before using any kind of reconstruction method, we need to determine the multiplicative constants. We have given two methods to determine these intensity values. The first is a heuristic method and the second method based on the consistency condition derived for absorbed projections [8].

We have successfully reconstructed the binary matrices from the absorption projections after determination of the intensity values.

Alkalmazott Matematikai Lapok **26** (2009), 351-363.

TERMELÉSI FÜGGVÉNYEK ÉS JELLEMZÉSEIK

NYUL BALÁZS

A közgazdaságtanban fontos szerepet játszó termelési függvények és ezek jellemzőinek ismertetése után a két legfontosabb termelési függvény, a Cobb–Douglas-típusú és az Arrow–Chenery–Minhas–Solow-típusú termelési függvény esetén határozzuk meg ezeket a jellemzőket. Majd kvázilineáris függvények, valamint kváziösszegek segítségével adunk jellemzési tételeket a CD-típusú és az ACMS-típusú termelési függvényekre. A tételek W. Eichhorntól [5], illetve F. Stehlingtől [13] származnak. Az eredeti bizonyításokat egyszerűsítjük, valamint a bennük található hiányosságokat javítjuk és pótoljuk. A bizonyításokban függvényegyenletek megoldására van szükségünk.

Kulcsszavak: termelési függvény, Cobb–Douglas-típusú termelési függvény, Arrow–Chenery–Minhas–Solow-típusú termelési függvény, kvázilineáris függvény, kváziösszeg, függvényegyenlet.

Mathematics Subject Classification 2000: 91B38, 39B22

1. Bevezetés

A termelési függvények fontos szerepet játszanak a közgazdaságtanban, ezen belül is elsősorban a mikroökonómiában. A termelési függvények közgazdaságtani értelemben azt adják meg, hogy a termeléshez szükséges termelési tényezők bizonyos mennyisége esetén – adott technológiai fejlettség mellett – mekkora lehet a termelés maximális kibocsátása.

A dolgozat első részében termelési függvényekkel foglalkozunk, matematikailag definiáljuk ezeket, majd a termelési függvények legfontosabb jellemzőit (átlagtermék, határtermék, parciális volumenrugalmasság, teljes volumenrugalmasság, technikai helyettesítési határráta, helyettesítési rugalmasság) tárgyaljuk, és leírjuk ezek közgazdaságtani jelentését is. Közelebbről a két legfontosabb termelési függvényt, a Cobb–Douglas-típusú és az Arrow–Chenery–Minhas–Solow-típusú termelési függvényt ismertetjük.

A dolgozat második részében a Cobb–Douglas-típusú és az Arrow–Chenery–Minhas–Solow-típusú termelési függvényekre vonatkozó jellemzési tételeket bizonyítunk be. Ezek bizonyításában alapvető szerepet játszanak a függvényegyenletek. A fenti típusú termelési függvényeket a kvázilineáris függvények és a kváziösszegek segítségével jellemezzük. Az eredeti bizonyítások *Wolfgang Eichhorn*tól [5], illetve

*Frank Stehling*től [13] származnak, melyek azonban kissé bonyolultak és néhány hiba, hiányosság is található bennük. A bizonyításokat több helyen leegyszerűsítjük, és kiküszöböljük ezeket a hibákat és hiányosságokat.

2. Termelési függvények

Ebben a részben a termelési függvényeket és azok legismertebb jellemzőit ismertetjük. Ezután bevezetjük a két legismertebb termelési függvénytípust, a Cobb–Douglas-típusú és az Arrow–Minhas–Chenery–Solow-típusú termelési függvényt, amelyekre kiszámoljuk a korábban definiált jellemzőket.

2.1. Termelési függvények tulajdonságai

A dolgozat további részében $n \geq 2$ egész szám.

Definíció. Egy $P :]0, +\infty[^n \longrightarrow]0, +\infty[$ függvényt **termelési függvény**nek nevezünk.

A termelési függvény közgazdaságtani értelemben azt adja meg, hogy az n darab termelési tényező rögzített mennyisége esetén, egységnyi idő alatt mennyi lehet a maximális kibocsátás. Megjegyezzük, hogy a szakirodalomban néha még további feltételeket is feltesznek a termelési függvény definíciójában. Termelési függvényekkel több könyv is foglalkozik, például [10], [12], [14], [15], [16].

Definíció. Legyen $\alpha \in \mathbb{R}$. Ekkor az $F :]0, +\infty[^n \longrightarrow]0, +\infty[$ függvényt **α-adfokú homogén**nek nevezzük, ha minden $(x_1, \ldots, x_n) \in]0, +\infty[^n$ és $\lambda > 0$ esetén $F(\lambda x_1, \ldots, \lambda x_n) = \lambda^\alpha F(x_1, \ldots, x_n)$. Ha $\alpha = 1$, akkor az F függvényt **homogén**nek nevezzük.

Differenciálható függvények esetén az α-adfokú homogén függvények jellemzését adja a közismert Euler-tétel.

2.1. TÉTEL. **(Euler)**
Legyen $F :]0, +\infty[^n \longrightarrow]0, +\infty[$ differenciálható függvény. Ekkor F akkor és csak akkor α-adfokú homogén függvény, ha minden $(x_1, \ldots, x_n) \in]0, +\infty[^n$ esetén

$$\partial_1 F(x_1, \ldots, x_n) \cdot x_1 + \ldots + \partial_n F(x_1, \ldots, x_n) \cdot x_n = \alpha \cdot F(x_1, \ldots, x_n).$$

A továbbiakban a termelési függvények néhány fontos jellemzőjét definiáljuk és megadjuk azok közgazdaságtani jelentését is.

Definíció. Legyen $P :]0, +\infty[^n \longrightarrow]0, +\infty[$ egy termelési függvény. Ekkor

$$AP_i(x_1, \ldots, x_n) = \frac{P(x_1, \ldots, x_n)}{x_i} \qquad ((x_1, \ldots, x_n) \in]0, +\infty[^n)$$

az i-**edik** ($i \in \{1, \ldots, n\}$) **termelési tényező átlagterméke** (average product).

Az i-edik termelési tényező átlagterméke azt adja meg, hogy ezen termelési tényező egy egységére az össztermelésből hány egység jut (a többi termelési tényező változatlansága mellett).

Definíció. Legyen $P :]0,+\infty[^n \longrightarrow]0,+\infty[$ egy differenciálható termelési függvény. Ekkor

$$MP_i(x_1,\ldots,x_n) = \partial_i P(x_1,\ldots,x_n) \qquad ((x_1,\ldots,x_n) \in]0,+\infty[^n)$$

az i-**edik** ($i \in \{1,\ldots,n\}$) **termelési tényező határterméke** (marginal product).

Az i-edik termelési tényező határterméke azt adja meg, hogy ezen termelési tényező 1 egységnyi növelésével (a többi termelési tényező változatlansága mellett) hány egységgel változik az összkibocsátás.

Definíció. Legyen $P :]0,+\infty[^n \longrightarrow]0,+\infty[$ egy differenciálható termelési függvény. Ekkor

$$\varepsilon_i(x_1,\ldots,x_n) = \frac{MP_i(x_1,\ldots,x_n)}{AP_i(x_1,\ldots,x_n)} \qquad ((x_1,\ldots,x_n) \in]0,+\infty[^n)$$

az i-**edik** ($i \in \{1,\ldots,n\}$) **termelési tényező parciális volumenrugalmassága** (partial elasticity of production).

Az i-edik termelési tényező parciális volumenrugalmassága azt adja meg, hogy ezen termelési tényező mennyiségét 1%-kal megnövelve (a többi termelési tényező változatlansága mellett), hány %-kal változik az összkibocsátás.

Definíció. Legyen $P :]0,+\infty[^n \longrightarrow]0,+\infty[$ egy differenciálható termelési függvény. Ekkor

$$\varepsilon(x_1,\ldots,x_n) = \varepsilon_1(x_1,\ldots,x_n) + \ldots + \varepsilon_n(x_1,\ldots,x_n) \quad ((x_1,\ldots,x_n) \in]0,+\infty[^n)$$

a P **termelési függvény teljes volumenrugalmassága** (total elasticity of production).

A termelési függvény teljes volumenrugalmassága azt adja meg, hogy minden termelési tényező mennyiségét 1%-kal megnövelve, körülbelül hány %-kal változik az összkibocsátás.

Euler tételéből következik, hogy egy α-adfokú homogén, differenciálható termelési függvény teljes volumenrugalmassága egyenlő a homogenitás fokával.

Definíció. Legyen $P :]0,+\infty[^n \longrightarrow]0,+\infty[$ egy differenciálható termelési függvény, melyre $\partial_k P(x_1,\ldots,x_n) \neq 0$ minden $k \in \{1,\ldots,n\}$ és $(x_1,\ldots,x_n) \in]0,+\infty[^n$ esetén. Ekkor

$$MRTS_{ij}(x_1,\ldots,x_n) = \frac{MP_i(x_1,\ldots,x_n)}{MP_j(x_1,\ldots,x_n)} \qquad ((x_1,\ldots,x_n) \in]0,+\infty[^n)$$

az i-**edik termelési tényező** j-**edik** ($i,j \in \{1,\ldots,n\}$) **termelési tényezőre vonatkozó technikai helyettesítési határrátája** (marginal rate of technical substitution).

Az i-edik termelési tényező j-edik termelési tényezőre vonatkozó technikai helyettesítési határrátа azt adja meg, hogy az i-edik termelési tényező mennyiségének 1 egységgel történő növelésével mennyivel kell csökkenteni a j-edik termelési tényező mennyiségét, hogy a megadott termelési szinten maradjunk.

Definíció. Legyen $P :]0,+\infty[^n \longrightarrow]0,+\infty[$ egy kétszer differenciálható termelési függvény, melyre $\partial_k P(x_1,\dots,x_n) \neq 0$ minden $k \in \{1,\dots,n\}$ és $(x_1,\dots,x_n) \in]0,+\infty[^n$ esetén. Ekkor feltéve, hogy a nevezőben levő kifejezés nem egyenlő 0-val,

$$ES_{ij}(x_1,\dots,x_n) = -\frac{\frac{1}{\partial_i P(x_1,\dots,x_n)\cdot x_i} + \frac{1}{\partial_j P(x_1,\dots,x_n)\cdot x_j}}{\frac{\partial_i^2 P(x_1,\dots,x_n)}{(\partial_i P(x_1,\dots,x_n))^2} - 2\frac{\partial_i\partial_j P(x_1,\dots,x_n)}{\partial_i P(x_1,\dots,x_n)\cdot\partial_j P(x_1,\dots,x_n)} + \frac{\partial_j^2 P(x_1,\dots,x_n)}{(\partial_j P(x_1,\dots,x_n))^2}}$$
$$((x_1,\dots,x_n) \in]0,+\infty[^n)$$

az i-edik termelési tényező j-edik $(i,j \in \{1,\dots,n\}, i \neq j)$ **termelési tényezőre vonatkozó helyettesítési rugalmassága** (elasticity of substitution).

Az i-edik termelési tényező j-edik termelési tényezőre vonatkozó helyettesítési rugalmassága azt adja meg, hogy adott termelési szinten az i-edik termelési tényező j-edik termelési tényezőre vonatkozó technikai helyettesítési határrátájának 1%-os növelésével, hány %-kal kell növelni a j-edik és az i-edik termelési tényezők felhasznált mennyiségének arányát.

Definíció. Egy $P :]0,+\infty[^n \longrightarrow]0,+\infty[$ kétszer differenciálható termelési függvény teljesíti a **CES-tulajdonság**ot (constant elasticity of substitution), ha minden $(x_1,\dots,x_n) \in]0,+\infty[^n$ esetén létezik $ES_{ij}(x_1,\dots,x_n)$, és valamilyen $c \in \mathbb{R}$-re $ES_{ij}(x_1,\dots,x_n) = c$ minden $i,j \in \{1,\dots,n\}$, $i \neq j$ esetén.

2.2. A Cobb–Douglas-típusú termelési függvény

Charles W. Cobb matematikus és *Paul H. Douglas* közgazdász több vizsgálatot végeztek arra vonatkozóan, hogy termelési függvények segítségével hogyan lehet leírni a nemzeti jövedelem megoszlását a munkás- és tőkésosztály között. Ennek eredményeként született meg 1928-ban a Cobb–Douglas-típusú termelési függvény [4].

Definíció. A $P :]0,+\infty[^n \longrightarrow]0,+\infty[$ termelési függvény **Cobb–Douglas-típusú (CD-típusú) termelési függvény**, ha

$$P(x_1,\dots,x_n) = Cx_1^{\alpha_1}\cdot\ldots\cdot x_n^{\alpha_n},$$

ahol $C > 0$, $\alpha_1 \neq 0,\dots,\alpha_n \neq 0$ olyan valós számok, hogy $\alpha = \sum_{i=1}^{n} \alpha_i \neq 0$.

2.2. TÉTEL. **(A CD-típusú termelési függvény tulajdonságai)**

(1) A CD-típusú termelési függvény α-adfokú homogén függvény.

A CD-típusú termelési függvényekre minden $(x_1,\dots,x_n)\in\,]0,+\infty[^n$ *esetén*

(2) $AP_i(x_1,\dots,x_n)=Cx_1^{\alpha_1}\cdot\ldots\cdot x_i^{\alpha_i-1}\cdot\ldots\cdot x_n^{\alpha_n}$ $(i\in\{1,\dots,n\})$,

(3) $MP_i(x_1,\dots,x_n)=C\alpha_i x_1^{\alpha_1}\cdot\ldots\cdot x_i^{\alpha_i-1}\cdot\ldots\cdot x_n^{\alpha_n}$ $(i\in\{1,\dots,n\})$,

(4) $\varepsilon_i(x_1,\dots,x_n)=\alpha_i$ $(i\in\{1,\dots,n\})$,

(5) $\varepsilon(x_1,\dots,x_n)=\alpha$,

(6) $MRTS_{ij}(x_1,\dots,x_n)=\dfrac{\alpha_i}{\alpha_j}\cdot\dfrac{x_j}{x_i}$ $(i,j\in\{1,\dots,n\})$,

(7) $ES_{ij}(x_1,\dots,x_n)=1$ $(i,j\in\{1,\dots,n\}, i\neq j)$, *így a CD-típusú termelési függvény teljesíti a CES-tulajdonságot.*

2.3. Az Arrow–Chenery–Minhas–Solow-típusú termelési függvény

Több bírálat is érte a Cobb–Douglas-típusú termelési függvény azon tulajdonságát, hogy helyettesítési rugalmassága 1. Ezt a tulajdonságot bírálták *Kenneth J. Arrow, Hollis B. Chenery, Bagicha S. Minhas* és *Robert M. Solow* [3] közgazdászok is. Számos ország több iparágát vizsgálták, és arra az eredményre jutottak, hogy a helyettesítési rugalmasság legtöbbször különbözik 1-től. Így 1961-ben egy új típusú termelési függvényt vezettek be, mely ugyan állandó helyettesítési rugalmasságú, de ezek a helyettesítési rugalmasságok a különböző iparágakban eltérőek.

Definíció. A $P:\,]0,+\infty[^n\longrightarrow\,]0,+\infty[$ termelési függvény **Arrow–Chenery–Minhas–Solow-típusú (ACMS-típusú) termelési függvény**, ha

$$P(x_1,\dots,x_n)=(\beta_1x_1^{-\varrho}+\ldots+\beta_nx_n^{-\varrho})^{-\frac{\alpha}{\varrho}},$$

ahol $\beta_1>0,\dots,\beta_n>0$, $\alpha\neq0$ és $\varrho\neq0$ valós számok.

A L'Hospital-szabály segítségével igazolható, hogy ha $\sum_{i=1}^{n}\beta_i=1$, akkor az ACMS-típusú termelési függvény $\varrho\to0$-val vett határértéke létezik és CD-típusú termelési függvény.

2.3. TÉTEL. **(Az ACMS-típusú termelési függvény tulajdonságai)**

(1) Az ACMS-típusú termelési függvény α-adfokú homogén függvény.

Az ACMS-típusú termelési függvényekre minden $(x_1,\dots,x_n)\in\,]0,+\infty[^n$ *esetén*

(2) $AP_i(x_1,\dots,x_n)=\dfrac{(\beta_1x_1^{-\varrho}+\ldots+\beta_ix_i^{-\varrho}+\ldots+\beta_nx_n^{-\varrho})^{-\frac{\alpha}{\varrho}}}{x_i}$ $(i\in\{1,\dots,n\})$,

(3) $MP_i(x_1,\dots,x_n)=\alpha(\beta_1x_1^{-\varrho}+\ldots+\beta_ix_i^{-\varrho}+\ldots+\beta_nx_n^{-\varrho})^{-\frac{\alpha}{\varrho}-1}\beta_ix_i^{-\varrho-1}$ $(i\in\{1,\dots,n\})$,

(4) $\varepsilon_i(x_1,\ldots,x_n) = \dfrac{\alpha\beta_i x_i^{-\varrho}}{\beta_1 x_1^{-\varrho} + \ldots + \beta_i x_i^{-\varrho} + \ldots + \beta_n x_n^{-\varrho}}$ $(i \in \{1,\ldots,n\})$,

(5) $\varepsilon(x_1,\ldots,x_n) = \alpha$,

(6) $MRTS_{ij}(x_1,\ldots,x_n) = \dfrac{\beta_i}{\beta_j} \cdot \left(\dfrac{x_j}{x_i}\right)^{1+\varrho}$ $(i,j \in \{1,\ldots,n\})$,

(7) *$\varrho \neq -1$ esetén* $ES_{ij}(x_1,\ldots,x_n) = \dfrac{1}{1+\varrho}$ $(i,j \in \{1,\ldots,n\}, i \neq j)$, *így ebben az esetben az ACMS-típusú termelési függvény teljesíti a CES-tulajdonságot.*

Megjegyzés. $\varrho = -1$ esetén a helyettesítési rugalmasság definíciójában szereplő tört nevezője 0, így ekkor az ACMS-típusú termelési függvény helyettesítési rugalmassága nincs értelmezve.

3. Jellemzési tételek

A továbbiakban először megoldunk egy függvényegyenletet, amire a későbbi jellemzési tételek bizonyításában lesz szükségünk. Ezután bevezetjük a kvázilinearitás fogalmát, majd definiáljuk a kváziösszeget, és ezek segítségével külön-külön jellemezzük a CD- és ACMS-típusú termelési függvényeket. Az előbbi jellemzési tétel *Wolfgang Eichhorn*tól [5], utóbbi pedig *Frank Stehling*től [13] származik. A két jellemzési tétel itt közölt bizonyítása egyszerűsíti, pontosítja és javítja a szerzők eredeti bizonyításait.

3.1. Egy függvényegyenlet megoldása

A következő tételben szereplő, később szükséges (1) függvényegyenlet megoldása megtalálható [2]-ben $(]0,1]$ értelmezési tartomány mellett), illetve [5]-ben. A teljesség kedvéért röviden megadjuk a tétel bizonyítását.

3.1. TÉTEL. *Legyen* $f :]0,+\infty[\longrightarrow \mathbb{R}$ *szigorúan monoton függvény, és legyenek* $r, q :]0,+\infty[\longrightarrow \mathbb{R}$ *függvények, melyekre teljesül, hogy minden* $\lambda, x \in]0,+\infty[$ *esetén*

$$f(\lambda x) = r(\lambda) f(x) + q(\lambda). \tag{1}$$

Ekkor léteznek olyan $\gamma \neq 0$ *és* δ *valós számok, hogy*

$$f(x) = \gamma \ln x + \delta, \quad r(\lambda) = 1, \quad q(\lambda) = \gamma \ln \lambda,$$

vagy léteznek olyan $\gamma \neq 0$, $\varrho \neq 0$ *és* δ *valós számok, hogy*

$$f(x) = \gamma x^{-\varrho} + \delta, \quad r(\lambda) = \lambda^{-\varrho}, \quad q(\lambda) = \delta(1 - \lambda^{-\varrho}).$$

Bizonyítás. Ha $x = 1$-et helyettesítünk az (1) függvényegyenletbe, majd ebből kivonjuk (1)-et, akkor azt kapjuk, hogy

$$f(\lambda) - f(\lambda x) = r(\lambda)(f(1) - f(x)).$$

Legyen $c = f(1)$ és $h :]0, +\infty[\longrightarrow \mathbb{R}$, $h(x) = f(x) - c$, ami szigorúan monoton függvény. Az előző egyenlet a most bevezetett jelölésekkel

$$h(\lambda x) = h(\lambda) + r(\lambda)h(x) \tag{2}$$

alakúra hozható. Ebben felcserélhetjük x és λ szerepét, amiből

$$h(x)(r(\lambda) - 1) = h(\lambda)(r(x) - 1) \tag{3}$$

adódik. A továbbiakban a bizonyítást két eset megkülönböztetésével folytatjuk.
1. eset: Ha $r(\lambda) = 1$ minden $\lambda \in]0, +\infty[$ esetén, akkor (2)-ből $h(\lambda x) = h(\lambda) + h(x)$ adódik. Jól ismert, hogy h szigorú monotonitása miatt $h(x) = \gamma \ln x$ $(x \in]0, +\infty[)$, ahol $\gamma \neq 0$ valós szám (lásd például [7] Chapter XIII. § 1., [6] Theorem 1.7.1.). Használjuk a $\delta = c$ jelölést. Ekkor $f(x) = \gamma \ln x + \delta$ és (1) alapján $q(\lambda) = \gamma \ln \lambda$.
2. eset: Ha létezik $\lambda_0 \in]0, +\infty[$, hogy $r(\lambda_0) \neq 1$, akkor λ_0-t behelyettesítve (3)-ba, $0 \neq r(\lambda_0) - 1$-gyel osztva, majd a $\gamma = \dfrac{h(\lambda_0)}{r(\lambda_0) - 1}$ jelölést használva

$$h(x) = \gamma(r(x) - 1) \tag{4}$$

adódik. Mivel f szigorúan monoton, ezért $\gamma \neq 0$. Ekkor (4)-et behelyettesítve (2)-be azt kapjuk, hogy $r(\lambda x) = r(\lambda)r(x)$. A (4) egyenlet miatt r szigorúan monoton függvény. Ekkor ugyancsak jól ismert, hogy $r(\lambda) = \lambda^{-\varrho}$ $(\lambda \in]0, +\infty[)$, ahol $\varrho \neq 0$ valós szám (lásd például [7] Chapter XIII. § 1., [6] Remark 1.9.23.). Legyen $\delta = -\gamma + c$, amivel ekkor $f(x) = \gamma x^{-\varrho} + \delta$. Mindezekből az (1) egyenletbe való visszahelyettesítéssel adódik, hogy $q(\lambda) = \delta(1 - \lambda^{-\varrho})$. □

3.2. Jellemzés kvázilinearitással

A kvázilineáris függvények már 1947-ben megjelentek [1] a kváziaritmetikai középértékek jellemzésével kapcsolatban.

Definíció. Egy $F :]0, +\infty[^n \longrightarrow]0, +\infty[$ függvény **kvázilineáris**, ha létezik $f :]0, +\infty[\longrightarrow \mathbb{R}$ folytonos, szigorúan monoton függvény, és léteznek olyan $a_1 \neq 0, \ldots, a_n \neq 0, b$ valós számok, hogy minden $(x_1, \ldots, x_n) \in]0, +\infty[^n$ esetén $a_1 f(x_1) + \ldots + a_n f(x_n) + b \in f(]0, +\infty[)$ és

$$F(x_1, \ldots, x_n) = f^{-1}(a_1 f(x_1) + \ldots + a_n f(x_n) + b).$$

3.2. TÉTEL. **(Eichhorn [5])** *Egy $F :]0, +\infty[^n \longrightarrow]0, +\infty[$ függvény akkor és csak akkor homogén és kvázilineáris, ha $\alpha = 1$ értékkel vett CD-típusú termelési függvény vagy ACMS-típusú termelési függvény.*

Bizonyítás.

I. Korábban már említettük, hogy az $\alpha = 1$ értékkel vett CD-típusú, illetve ACMS-típusú termelési függvények homogének. Továbbá a CD-típusú termelési függvény kvázilineáris az $f :]0, +\infty[\longrightarrow \mathbb{R}$, $f(x) = \ln x$, $a_i = \alpha_i$ $(i \in \{1, \ldots, n\})$, $b = \ln C$ választással. Hasonlóan az $\alpha = 1$ esetben az ACMS-típusú termelési függvény is kvázilineáris, méghozzá $f :]0, +\infty[\longrightarrow \mathbb{R}$, $f(x) = x^{-\varrho}$, $a_i = \beta_i$ $(i \in \{1, \ldots, n\})$, $b = 0$-val.

II. Legyen F kvázilineáris függvény a definícióbeli jelölésekkel. Mivel F homogén, így f alkalmazásával azt kapjuk, hogy minden $\lambda > 0$ és minden $(x_1, \ldots, x_n) \in]0, +\infty[^n$ esetén

$$a_1 f(\lambda x_1) + \ldots + a_n f(\lambda x_n) + b = f(\lambda f^{-1}(a_1 f(x_1) + \ldots + a_n f(x_n) + b)).$$

Adott $\lambda > 0$ mellett legyen $F_\lambda : f(]0, +\infty[) \longrightarrow \mathbb{R}$, $F_\lambda(y) = f(\lambda f^{-1}(y))$, és legyen $y_i = f(x_i)$ $(i \in \{1, \ldots, n\})$. Ekkor

$$a_1 F_\lambda(y_1) + \ldots + a_n F_\lambda(y_n) + b = F_\lambda(a_1 y_1 + \ldots + a_n y_n + b). \tag{5}$$

Legyenek $s_1, s_2 \in f(]0, +\infty[)$ rögzítettek, ahol $s_1 \neq s_2$. Az egyenlet mindkét oldalát y_1 szerint integrálva s_1-től s_2-ig

$$\int_{s_1}^{s_2} F_\lambda(a_1 y_1 + \ldots + a_n y_n + b)\, dy_1 =$$

$$= a_1 \int_{s_1}^{s_2} F_\lambda(y_1)\, dy_1 + (s_2 - s_1)(a_2 F_\lambda(y_2) + \ldots + a_n F_\lambda(y_n) + b)$$

adódik. A baloldalon elvégezve a $t = a_1 y_1 + \ldots + a_n y_n + b$ helyettesítést, átrendezés után következik, hogy

$$\int_{a_1 s_1 + a_2 y_2 + \ldots + a_n y_n + b}^{a_1 s_2 + a_2 y_2 + \ldots + a_n y_n + b} F_\lambda(t) \frac{1}{a_1}\, dt - a_1 \int_{s_1}^{s_2} F_\lambda(y_1)\, dy_1 -$$

$$- (s_2 - s_1)(a_3 F_\lambda(y_3) + \ldots + a_n F_\lambda(y_n) + b) = (s_2 - s_1) a_2 F_\lambda(y_2).$$

Mivel F_λ folytonos, így $y_2 \mapsto \int_{a_1 s_1 + a_2 y_2 + \ldots + a_n y_n + b}^{a_1 s_2 + a_2 y_2 + \ldots + a_n y_n + b} F_\lambda(t) \frac{1}{a_1}\, dt$ differenciálható, ezért F_λ differenciálható. Az (5) egyenlet mindkét oldalát y_1 szerint deriválva kapjuk, hogy $F'_\lambda(y_1) = F'_\lambda(a_1 y_1 + \ldots + a_n y_n + b)$. Hasonlóan deriválhatjuk y_2 szerint is, amiből $F'_\lambda(y_1) = F'_\lambda(y_2)$, azaz F'_λ konstans függvény. Így létezik $r, q \in \mathbb{R}$, hogy $F_\lambda(y) = ry + q$. Eddig rögzített λ-val dolgoztunk, amitől azonban az r és q értékei függhetnek, ezért $F_\lambda(y) = r(\lambda)y + q(\lambda)$, ahol $r, q :]0, +\infty[\longrightarrow \mathbb{R}$ függvények. Legyen $x = f^{-1}(y)$, amivel F_λ definíciójából $f(\lambda x) = r(\lambda) f(x) + q(\lambda)$. Ekkor a

3.1. tételből a következő két eset adódik.

1. eset: $f(x) = \gamma \ln x + \delta$, ahol $\gamma \neq 0$, δ valós számok. Ekkor $f^{-1}(y) = e^{\frac{y-\delta}{\gamma}}$. Ebben az esetben F egy CD-típusú termelési függvény, ahol $C = e^{\frac{a_1\delta+\ldots+a_n\delta+b-\delta}{\gamma}}$ és $\alpha_1 = a_1, \ldots, \alpha_n = a_n$. Mivel F homogén CD-típusú termelési függvény, így még $\alpha = \alpha_1 + \ldots + \alpha_n = 1$-nek is teljesülnie kell, amiből $C = e^{\frac{b}{\gamma}}$.
2. eset: $f(x) = \gamma x^{-\varrho} + \delta$, ahol $\gamma \neq 0$, $\varrho \neq 0$, δ valós számok. Ebben az esetben $f^{-1}(y) = \left(\frac{y-\delta}{\gamma}\right)^{-\frac{1}{\varrho}}$.

Ekkor belátjuk, hogy $a_i > 0$ $(i \in \{1, \ldots, n\})$. Nyilvánvaló, hogy $\gamma > 0$ esetén $f(]0, +\infty[) =]\delta, +\infty[$, míg $\gamma < 0$ esetén $f(]0, +\infty[) =]-\infty, \delta[$. A kvázilinearitás miatt minden $(x_1, \ldots, x_n) \in]0, +\infty[^n$ esetén

$$a_1(\gamma x_1^{-\varrho} + \delta) + \ldots + a_n(\gamma x_n^{-\varrho} + \delta) + b \in f(]0, +\infty[).$$

Indirekt tegyük fel, hogy például $a_1 < 0$, az $x_2, \ldots, x_n$ értékét pedig tetszőlegesen rögzítsük le. Tartsunk x_1-gyel 0-hoz vagy $+\infty$-hez úgy, hogy $x_1^{-\varrho}$ határértéke $+\infty$ legyen. Ekkor az előző kifejezés határértéke $\gamma > 0$ esetén $-\infty$, $\gamma < 0$ esetén pedig $+\infty$, ami ellentmond a fentieknek.
Ebben az esetben

$$F(x_1, \ldots, x_n) = \left(a_1 x_1^{-\varrho} + \ldots + a_n x_n^{-\varrho} + \frac{a_1\delta + \ldots + a_n\delta + b - \delta}{\gamma}\right)^{-\frac{1}{\varrho}}.$$

Vezessük be a továbbiakban a $Q = \frac{a_1\delta+\ldots+a_n\delta+b-\delta}{\gamma}$ jelölést. Mivel F homogén, így minden $\lambda > 0$ esetén teljesülnie kell, hogy

$$\left(a_1(\lambda x_1)^{-\varrho} + \ldots + a_n(\lambda x_n)^{-\varrho} + Q\right)^{-\frac{1}{\varrho}} = \lambda\left(a_1 x_1^{-\varrho} + \ldots + a_n x_n^{-\varrho} + Q\right)^{-\frac{1}{\varrho}}.$$

Az egyenlet mindkét oldalának $-\varrho$-adik hatványra emelése után $(1 - \lambda^{-\varrho})Q = 0$ kapható. Mivel $\varrho \neq 0$ és minden $\lambda > 0$-ra teljesül ez az egyenlet, így $Q = 0$, tehát ekkor F egy ACMS-típusú termelési függvény $\alpha = 1$-gyel és $\beta_1 = a_1, \ldots, \beta_n = a_n$-nel, amelyekről már beláttuk, hogy pozitívak. □

3.3. Jellemzés kváziösszegekkel

A kéttagú kváziösszeg fogalma *Aczél János*tól származik, többtagú kváziösszegek is előfordulnak az irodalomban, például a konzisztens aggregációval kapcsolatban [8], [9].

Definíció. Egy $F :]0, +\infty[^n \longrightarrow]0, +\infty[$ függvény **kváziösszeg**, ha léteznek $g_i :]0, +\infty[\longrightarrow \mathbb{R}$ $(i \in \{1, \ldots, n\})$ folytonos, szigorúan monoton függvények, és létezik $I \subseteq \mathbb{R}$ pozitív hosszúságú intervallum, $g : I \longrightarrow]0, +\infty[$ folytonos, szigorúan monoton függvény, hogy minden $(x_1, \ldots, x_n) \in]0, +\infty[^n$ esetén

$g_1(x_1) + \ldots + g_n(x_n) \in I$ és

$$F(x_1, \ldots, x_n) = g(g_1(x_1) + \ldots + g_n(x_n)).$$

A kváziösszeg függvényeket szokták általánosított kvázilineáris függvényeknek is nevezni a következő állítás miatt.

ÁLLÍTÁS. *Minden kvázilineáris függvény kváziösszeg.*

Bizonyítás. Legyen F egy kvázilineáris függvény a korábbi jelölések megtartásával. Legyen ekkor $g_i :]0, +\infty[\longrightarrow \mathbb{R}$, $g_i(x) = a_i f(x)$ $(i \in \{1, \ldots, n-1\})$,

$$g_n :]0, +\infty[\longrightarrow \mathbb{R}, \quad g_n(x) = a_n f(x) + b,$$
$$g : f(]0, +\infty[) \longrightarrow]0, +\infty[, \quad g(x) = f^{-1}(x).$$

□

3.3. TÉTEL. **(Stehling [13])** *Egy $F :]0, +\infty[^n \longrightarrow]0, +\infty[$ függvény akkor és csak akkor α-adfokú homogén ($\alpha \neq 0$) és kváziösszeg, ha CD-típusú termelési függvény vagy ACMS-típusú termelési függvény.*

Bizonyítás.

I. A korábban mondottak szerint a CD-típusú, illetve az ACMS-típusú termelési függvények α-adfokú homogén függvények. A CD-típusú termelési függvény kváziösszeg

$$g_i :]0, +\infty[\longrightarrow \mathbb{R}, \quad g_i(x) = \alpha_i \ln x \quad (i \in \{1, \ldots, n-1\}),$$
$$g_n :]0, +\infty[\longrightarrow \mathbb{R}, \quad g_n(x) = \alpha_n \ln x + \ln C,$$

továbbá $g : \mathbb{R} \longrightarrow]0, +\infty[$, $g(x) = e^x$-szel. Továbbá az ACMS-típusú termelési függvény is kváziösszeg $g_i :]0, +\infty[\longrightarrow \mathbb{R}$, $g_i(x) = \beta_i x^{-\varrho}$ $(i \in \{1, \ldots, n\})$, továbbá $g :]0, +\infty[\longrightarrow]0, +\infty[$, $g(x) = x^{-\frac{u}{v}}$ választással.

II. Legyen F kváziösszeg a definícióbeli jelölésekkel. Mivel F α-adfokú homogén függvény, így minden $\lambda > 0$ és $(x_1, \ldots, x_n) \in]0, +\infty[^n$ esetén

$$g(g_1(\lambda x_1) + \ldots + g_n(\lambda x_n)) = \lambda^\alpha g(g_1(x_1) + \ldots + g_n(x_n)). \tag{6}$$

Az egyenlet mindkét oldalára alkalmazzuk g^{-1}-et, és legyen

$$y_i = g_i(x_i) \quad (i \in \{1, \ldots, n\}),$$

továbbá adott $\lambda > 0$ mellett legyen $H_\lambda : I \longrightarrow \mathbb{R}$, $H_\lambda(y) = g^{-1}(\lambda^\alpha g(y))$ és $h_\lambda^{(i)} : g_i(]0, +\infty[) \longrightarrow \mathbb{R}$, $h_\lambda^{(i)}(y) = g_i(\lambda g_i^{-1}(y))$ $(i \in \{1, \ldots, n\})$. Ekkor azt kapjuk, hogy

$$h_\lambda^{(1)}(y_1) + \ldots + h_\lambda^{(n)}(y_n) = H_\lambda(y_1 + \ldots + y_n). \tag{7}$$

Legyenek $s_1, s_2 \in g_1(]0, +\infty[)$ rögzítettek, ahol $s_1 \neq s_2$. Az egyenlet mindkét oldalát y_1 szerint s_1-től s_2-ig integrálva

$$\int_{s_1}^{s_2} H_\lambda(y_1 + \ldots + y_n)\, dy_1 = \int_{s_1}^{s_2} h_\lambda^{(1)}(y_1)\, dy_1 + (s_2 - s_1)(h_\lambda^{(2)}(y_2) + \ldots + h_\lambda^{(n)}(y_n))$$

adódik. A baloldalon elvégezve a $t = y_1 + \ldots + y_n$ helyettesítést és átrendezve az egyenletet

$$\int_{s_1+y_2+\ldots+y_n}^{s_2+y_2+\ldots+y_n} H_\lambda(t)\,dt - \int_{s_1}^{s_2} h_\lambda^{(1)}(y_1)\,dy_1 - (s_2 - s_1)(h_\lambda^{(3)}(y_3) + \ldots + h_\lambda^{(n)}(y_n)) =$$

$$= (s_2 - s_1)h_\lambda^{(2)}(y_2)$$

következik. Mivel H_λ folytonos, így $y_2 \mapsto \int_{s_1+y_2+\ldots+y_n}^{s_2+y_2+\ldots+y_n} H_\lambda(t)\,dt$ differenciálható. Ekkor az előző egyenlet szerint $h_\lambda^{(2)}$ differenciálható. Hasonlóan $h_\lambda^{(i)}$ $(i \in \{1,\ldots,n\})$ differenciálható, így (7) miatt $H_\lambda|\ g_1(]0,+\infty[)+\ldots+g_n(]0,+\infty[)$ is differenciálható. A (7) egyenlet mindkét oldalát y_1 szerint deriválva $H'_\lambda(y_1 + \ldots + y_n) = h_\lambda^{(1)'}(y_1)$ adódik. Hasonlóan deriválhatunk y_2 szerint is, amiből $h_\lambda^{(1)'}(y_1) = h_\lambda^{(2)'}(y_2)$ következik, azaz $h_\lambda^{(1)'}$ és $h_\lambda^{(2)'}$ egyenlő értékű konstans függvények. Hasonlóan adódik, hogy $h_\lambda^{(i)'}$ $(i \in \{1,\ldots,n\})$ páronként egyenlő értékű konstans függvények. Tehát léteznek olyan $r, q_1, \ldots, q_n$ valós számok, hogy $h_\lambda^{(i)}(y) = ry + q_i$ $(i \in \{1,\ldots,n\})$. Eddig rögzített λ-val dolgoztunk, amelytől azonban az $r, q_1, \ldots, q_n$ értékei függhetnek, amiből azt kapjuk, hogy $h_\lambda^{(i)}(y) = r(\lambda)y + q_i(\lambda)$ $(i \in \{1,\ldots,n\})$, ahol $r, q_1, \ldots, q_n :]0,+\infty[\longrightarrow \mathbb{R}$ függvények. Legyen $x = g_i^{-1}(y)$. Ekkor $h_\lambda^{(i)}$ definíciójából kapjuk, hogy $g_i(\lambda x) = r(\lambda)g_i(x) + q_i(\lambda)$ $(i \in \{1,\ldots,n\})$. Mivel minden $i \in \{1,\ldots,n\}$ esetén az r függvény közös, így a 3.1. tételből az alábbi két eset adódik.

1. eset: $g_i(x) = \gamma_i \ln x + \delta_i$, ahol $\gamma_i \neq 0$, δ_i valós számok $(i \in \{1,\ldots,n\})$. Ekkor $g_i(]0,+\infty[) = \mathbb{R}$ $(i \in \{1,\ldots,n\})$, ezért ebben az esetben $I = \mathbb{R}$. Vezessük be a $\delta = \delta_1 + \ldots + \delta_n$ és $\gamma = \gamma_1 + \ldots + \gamma_n$ jelöléseket. Ekkor a (6) egyenlet alapján

$$g(\gamma_1 \ln x_1 + \ldots + \gamma_n \ln x_n + \gamma \ln \lambda + \delta) = \lambda^\alpha g(\gamma_1 \ln x_1 + \ldots + \gamma_n \ln x_n + \delta)$$

adódik. Ebbe $x_1 = e^{\frac{y-\delta}{\gamma_1}}, x_2 = \ldots = x_n = 1$-et helyettesítve

$$g(y + \gamma \ln \lambda) = \lambda^\alpha g(y) \tag{8}$$

következik. Mivel $g(y) > 0$ és $\alpha \neq 0$, így $\gamma \neq 0$. Helyettesítsünk $y = 0$-t (8)-ba, és használjuk a $d = g(0) > 0$ és az $u = \gamma \ln \lambda$ jelöléseket. Ebből $g(u) = d \cdot e^{\frac{\alpha}{\gamma}u}$. Ekkor tehát F egy CD-típusú termelési függvény, ahol $C = d \cdot e^{\frac{\alpha}{\gamma}\delta} > 0$, $\alpha_i = \frac{\alpha}{\gamma}\gamma_i \neq 0$ $(i \in \{1,\ldots,n\})$, és $\sum_{i=1}^{n} \alpha_i = \sum_{i=1}^{n} \frac{\alpha}{\gamma}\gamma_i = \alpha \neq 0$.

2. eset: $g_i(x) = \gamma_i x^{-\varrho} + \delta_i$, ahol $\gamma_i \neq 0$, $\varrho \neq 0$, δ_i valós számok $(i \in \{1, \ldots, n\})$. Vezessük be a $\delta = \delta_1 + \ldots + \delta_n$ és $\gamma = \gamma_1 + \ldots + \gamma_n$ jelöléseket. Ekkor (6)-ba helyettesítve kapjuk, hogy

$$g(\gamma_1 \cdot (\lambda x_1)^{-\varrho} + \ldots + \gamma_n \cdot (\lambda x_n)^{-\varrho} + \delta) = \lambda^\alpha g(\gamma_1 x_1^{-\varrho} + \ldots + \gamma_n x_n^{-\varrho} + \delta). \quad (9)$$

Belátjuk, hogy $\gamma_1, \ldots, \gamma_n$ azonos előjelűek. Indirekt tegyük fel, hogy létezik közöttük pozitív és negatív is. Ekkor $v = \gamma_1 x_1^{-\varrho} + \ldots + \gamma_n x_n^{-\varrho}$ minden valós számot felvesz értékként. Ezzel a jelöléssel $g(\lambda^{-\varrho} v + \delta) = \lambda^\alpha g(v + \delta)$, amibe $v = 0$-t írva $g(\delta) > 0$ miatt ellentmondáshoz jutunk.
Ekkor ha $\gamma_i > 0$, akkor $g_i(]0, +\infty[) =]\delta_i, +\infty[$ $(i \in \{1, \ldots, n\})$, ezért $]\delta, +\infty[\subseteq I$.
Ha pedig $\gamma_i < 0$, akkor $g_i(]0, +\infty[) =]-\infty, \delta_i[$ $(i \in \{1, \ldots, n\})$, ezért $]-\infty, \delta[\subseteq I$.

Helyettesítsünk (9)-be $x_1 = \ldots = x_n = 1$-et. Ekkor a

$$g(\gamma\lambda^{-\varrho} + \delta) = \lambda^\alpha g(\gamma + \delta) \quad (10)$$

egyenlethez jutunk. Az előzőekhez hasonlóan kapjuk, hogy $\gamma \neq 0$.

Használjuk a $d = g(\gamma + \delta) > 0$ és az $u = \gamma\lambda^{-\varrho} + \delta$ jelöléseket. Az u a γ_i-k előjelétől függően bármilyen értéket felvehet $]\delta, +\infty[$ vagy $]-\infty, \delta[$-ból. Ezekkel (10)-ből $g(u) = d\left(\frac{u-\delta}{\gamma}\right)^{-\frac{\alpha}{\varrho}}$ adódik. Ekkor kihasználva, hogy $d > 0$ és $\alpha \neq 0$, következik, hogy F egy ACMS-típusú termelési függvény $\beta_i = d^{-\frac{\varrho}{\alpha}} \frac{\gamma_i}{\gamma}$-val $(i \in \{1, \ldots, n\})$, amelyek a korábban belátottak szerint pozitívak. □

Hivatkozások

[1] J. Aczél: *On mean values*, Bulletin of the American Mathematical Society **54** (1948), 392–400.

[2] J. Aczél, Z. Daróczy: *On Measures of Information and Their Characterizations*, Academic Press, 1975.

[3] K. J. Arrow, H. B. Chenery, B. S. Minhas, R. M. Solow: *Capital-labor substitution and economic efficiency*, The Review of Economics and Statistics **43** (1961), 225–250.

[4] C. W. Cobb, P. H. Douglas: *A theory of production*, The American Economic Review **18** (1928), 139–165.

[5] W. Eichhorn: *Characterization of the CES production functions by quasilinearity*, in: Production Theory (W. Eichhorn, R. Henn, O. Opitz and R. W. Shephard eds.), Springer-Verlag, 1974, 21–33.

[6] W. Eichhorn: *Functional Equations in Economics*, Addison-Wesley Publishing Company, 1978.

[7] M. Kuczma: *An Introduction to the Theory of Functional Equations and Inequalities*, Państwowe Wydawnictwo Naukowe, 1985.

[8] GY. MAKSA: *Solution of generalized bisymmetry type equations without surjectivity assumptions*, Aequationes Mathematicae **57** (1999), 50–74.

[9] GY. MAKSA, E. NIZSALÓCZKI: *Quasi-sums in several variables*, Acta Mathematica Academiae Paedagogicae Nyíregyháziensis **22** (2006), 193–207.

[10] N. GREGORY MANKIW: *Makroökonómia*, Osiris Kiadó, 2005.

[11] MÁTYÁS A.: *A modern közgazdaságtan története*, Aula Kiadó, 2003.

[12] P. A. SAMUELSON, W. D. NORDHAUS: *Közgazdaságtan*, KJK-KERSZÖV Jogi és Üzleti Kiadó, 2003.

[13] F. STEHLING: *Eine neue Charakterisierung der CD- und ACMS-Produktionsfunktionen*, Operations Research-Verfahren **21** (1975), 222–238.

[14] K. SYDSÆTER, P. I. HAMMOND: *Matematika közgazdászoknak*, Aula Kiadó, 2006.

[15] H. R. VARIAN: *Mikroökonómia középfokon*, Akadémiai Kiadó, 2005.

[16] ZALAI E.: *Matematikai közgazdaságtan*, KJK-KERSZÖV Jogi és Üzleti Kiadó, 2000.

(Beérkezett: 2008. május 27.)

NYUL BALÁZS
Debreceni Egyetem
Matematikai Intézet
4010 Debrecen, Pf. 12.
nyulbalazs@unideb.hu

PRODUCTION FUNCTIONS AND THEIR CHARACTERIZATIONS

BALÁZS NYUL

We describe production functions that play an important role in economics, and define some properties of them. We calculate these values for production functions of Cobb-Douglas type and Arrow-Chenery-Minhas-Solow type. Then we give characterization theorems of production functions of CD type and ACMS type using the notion of quasilinear functions and quasisums. The theorems are due to W. Eichhorn [5] and F. Stehling [13]. We simplify the original proofs and correct the defects of them. In the proofs we need to solve functional equations.

Keywords: production function, production function of Cobb-Douglas type, production function of Arrow-Chenery-Minhas-Solow type, quasilinear function, quasisum, functional equation
Mathematics Subject Classification 2000: 91B38, 39B22

Alkalmazott Matematikai Lapok **26** (2009), 365-379.

ITERÁCIÓFÜGGETLEN LÉPÉSHOSSZ ÉS LÉPÉSBECSLÉS A DIKIN-ALGORITMUS ALKALMAZÁSÁBAN A LINEÁRIS PROGRAMOZÁSI FELADATRA

MIKLÓS ZOLTÁN, TAKÁCS SZABOLCS

Az alábbi cikkben bemutatjuk a Dikin ellipszoid módszert a lineáris programozási feladatra az [1] könyvben található fejezet alapján. A fenti hivatkozáson a komplexitás vizsgálat egyik bizonyítása hibás.

A hibát kijavítva az eredetinél pontosabb tételt kapunk. Egy iterációfüggetlen lépéshosszt határozunk meg a Dikin ellipszoid módszerhez, melynek segítségével az algoritmus lépésigényére felső becslést lehet adni.

A cikkben található jelölések a könnyebb érthetőség kedvéért az [1] könyv jelöléseivel egyeznek meg.

1. Bevezetés

Legyen $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^m$ és $\mathbf{c} \in \mathbb{R}^n$. Tekintsük az alábbi primál-duál feladatpárt:

$$\begin{aligned} &(\mathbf{P}) \qquad \min\left\{\mathbf{c}^T\mathbf{x} : A\mathbf{x} \geq \mathbf{b},\ \mathbf{x} \geq \mathbf{0}\right\}, \\ &(\mathbf{D}) \qquad \max\left\{\mathbf{b}^T\mathbf{y} : A^T\mathbf{y} \leq \mathbf{c},\ \mathbf{y} \geq \mathbf{0}\right\}. \end{aligned}$$

A gyenge dualitás tétel alkalmazásával az alábbi feltételek adódnak:

$$\begin{aligned} A\mathbf{x} - \mathbf{z} &= \mathbf{b}, & \mathbf{x} \geq \mathbf{0},\ \mathbf{z} \geq \mathbf{0}, \\ A^T\mathbf{y} + \mathbf{w} &= \mathbf{c}, & \mathbf{y} \geq \mathbf{0},\ \mathbf{w} \geq \mathbf{0}, \\ \mathbf{b}^T\mathbf{y} - \mathbf{c}^T\mathbf{x} &= \zeta, & \zeta \geq 0. \end{aligned}$$

Ezt homogenizálva a *Goldman–Tucker-rendszer*t kapjuk:

$$\begin{aligned} A\mathbf{x} - \mathbf{z} &= \xi\mathbf{b}, & \mathbf{x} \geq \mathbf{0},\ \mathbf{z} \geq \mathbf{0}, \\ -A^T\mathbf{y} - \mathbf{w} &= -\xi\mathbf{c}, & \mathbf{y} \geq \mathbf{0},\ \mathbf{w} \geq \mathbf{0}, \qquad (GT) \\ \mathbf{b}^T\mathbf{y} - \mathbf{c}^T\mathbf{x} &= \zeta, & \xi \geq 0,\ \zeta \geq 0. \end{aligned}$$

Vezessük be a fenti (GT) rendszerre az alábbi jelöléseket:

$$M = \begin{pmatrix} 0 & A & -\mathbf{b} \\ -A^T & 0 & \mathbf{c} \\ \mathbf{b}^T & -\mathbf{c}^T & 0 \end{pmatrix}, \qquad \mathbf{u} = \begin{pmatrix} \mathbf{y} \\ \mathbf{x} \\ \xi \end{pmatrix}, \qquad \mathbf{s} = \begin{pmatrix} \mathbf{z} \\ \mathbf{w} \\ \zeta \end{pmatrix} \qquad (GT).$$

Így az alábbi speciális ferdén szimmetrikus feladatot kell megoldanunk – ahol M ferdén szimmetrikus mátrix $\left(M^T = -M\right)$:

$$M\mathbf{u} = \mathbf{s}, \qquad \mathbf{u} \geq \mathbf{0},\ \mathbf{s} \geq \mathbf{0}.$$

A fenti rendszerre az alábbi tétel fogalmazható meg:

1.1. TÉTEL. *Legyen adott a (P) és (D) primál-duál lineáris programozási feladatpár. Ekkor:*

i. *A (P) és (D) feladatok tetszőleges* $(\mathbf{x}, \mathbf{y})$ *optimális megoldáspárja, a megfelelő (GT) rendszer egy megoldását adja* $\xi = 1,\ \zeta = 0$ *választással.*

ii. ***(Goldman–Tucker-tétel)*** *A (GT) egyenlőtlenség-rendszernek van szigorúan komplementáris megoldása, azaz olyan megoldás, melyre* $\mathbf{u} + \mathbf{s} > \mathbf{0}$.

iii. *A (GT) rendszer tetszőleges* $(\mathbf{u}, \mathbf{s}) = (\mathbf{y}, \mathbf{x}, \xi, \mathbf{z}, \mathbf{w}, \zeta)$ *megoldására:*

   (1) ***(Erős dualitás tétel)*** *Ha* $\xi > 0$ *és* $\zeta = 0$, *akkor* $\left(\frac{\mathbf{x}}{\xi}, \frac{\mathbf{y}}{\xi}\right)$ *optimális megoldáspárját adja a (P) és (D) feladatoknak.*

   (2) ***(Iránymenti korlátosság tétel)*** *Ha* $\xi = 0$ *és* $\zeta > 0$, *akkor vagy (P), vagy (D), vagy mindkettő nem megengedett. Amennyiben csak az egyik feladat üres, úgy a másik feladat nem korlátos.*

iv. *A* $\xi\zeta > 0$ *eset nem fordul elő.*

Az 1.1. tétel alapján elegendő a (GT) rendszert vizsgálni. A (GT) rendszer átírható az alábbi lineáris programozási feladattá:

$$\min\left\{\mathbf{0}^T\mathbf{u} : M\mathbf{u} = \mathbf{s},\ \mathbf{u} \geq \mathbf{0},\ \mathbf{s} \geq \mathbf{0}\right\}.$$

1.1. A ferdén szimmetrikus feladat

Általánosabb alakban szokás a (GT) feladatot felírni:

1.1. Definíció. $\mathbf{M} \in \mathbb{R}^{n\times n}$, $\mathbf{M}^T = -\mathbf{M}$, $\mathbf{q} \in \mathbb{R}^n$, $\mathbf{q} \geq 0$.

A ferdén szimmetrikus önduális feladat (Skew Symmetric Problem) – (SP):

$$\min\left\{\mathbf{x}^T\mathbf{s} : M\mathbf{x} + \mathbf{q} = \mathbf{s},\ \mathbf{x}, \mathbf{s} \geq 0\right\}.$$

A megengedettségi tartományt

$$\mathcal{F} = \left\{(\mathbf{x}, \mathbf{s}) \in \mathbb{R}^{2n} : M\mathbf{x} + \mathbf{q} = \mathbf{s},\ \mathbf{x} \geq 0,\ \mathbf{s} \geq 0\right\}\text{-el},$$

a belső pontok halmazát

$$\mathcal{F}^0 = \{(\mathbf{x},\mathbf{s}) \in \mathbb{R}^{2n} : M\mathbf{x}+\mathbf{q} = \mathbf{s},\ \mathbf{x} > 0,\ \mathbf{s} > 0\}\text{-al}$$

jelöljük.

Az (SP) feladatban $\mathbf{x}^T\mathbf{s} = \mathbf{q}^T\mathbf{x}$, mert bármely $\mathbf{x}$ vektorra $\mathbf{x}^T M\mathbf{x} = 0$ az M mátrix ferde szimmetriája miatt.

Tekintsük az alábbi definíciót:

1.2. Definíció. Az $\mathbf{x}$ pont *ε-optimális*, ha a dualitásrés legfeljebb ε, azaz ha $\mathbf{x}^T\mathbf{s} \le \varepsilon$.

$(SP)_\varepsilon$-nak nevezzük a következő megengedettségi feladatot:

$$(SP)_\varepsilon : \begin{cases} M\mathbf{x}+\mathbf{q} = \mathbf{s} & \text{(megoldás)} \\ \mathbf{x} > \mathbf{0},\ \mathbf{s} > \mathbf{0} & \text{(megengedett, sőt: } \textit{belső pont}\text{)} \\ \mathbf{x}^T\mathbf{s} \le \varepsilon & (\varepsilon\text{-optimális}). \end{cases}$$

Az ε-optimális belsőpontos megoldások halmazát $\mathcal{F}_\varepsilon^o$-al jelöljük.

Ismert egy kerekítési eljárás, mely egy megfelelő ε-optimális megoldásból elő tud állítani egy optimális megoldást – ehhez ε-t kellően kicsire kell választani. Erről bővebben [1] alatt található információ.

Elegendő tehát egy $(\mathbf{x},\mathbf{s}) \in \mathbb{R}^{2n}$ ε-optimális belső pontos megoldást előállítani.

1.3. Definíció. Legyen $\mu > 0$ tetszőleges. Ekkor az $(\mathbf{x},\mathbf{s}) > \mathbf{0}$ belső pontos megoldást, ha $\mathbf{xs} = \mu\mathbf{e}$, *$\mu$-centrum*nak nevezzük. A μ-centrumok összességét *centrális út*nak hívjuk és C-vel jelöljük.

Amennyiben létezik belső pont, úgy a μ-centrum egyértelműen létezik. Ha μ-vel tartunk a nullához belső pontok egy konvergens sorozata mentén, akkor:

i. a μ-centrumok határértéke szigorúan komplementáris megoldását adja az (SP) feladatnak

ii. ha a dualitásrés ε-t már nem haladja meg, úgy a μ-centrum az $(SP)_\varepsilon$ egy megoldását szolgáltatja.

Általános esetben elméletileg az ellipszoid módszerrel eldönthető, hogy létezik-e belső pont, vagy sem. Továbbá belső pontot is szolgáltat a végső ellipszoid középpontjaként. Ez a gyakorlatban nem praktikus.

Amennyiben $\mathbf{q} = 0$ – mint az eredeti esetben – úgy az alábbi beágyazás végezhető el:

$$\begin{aligned} \min \lambda\theta & \\ M\mathbf{x}+\mathbf{r}\theta &= \mathbf{s} \\ -\mathbf{r}^T\mathbf{x}+\lambda &= \nu \\ \mathbf{x},\theta,\mathbf{s},\nu &\ge 0, \end{aligned}$$

ahol $\mathbf{r} \in \mathbb{R}^n$ és $\lambda \in \mathbb{R}_+$. Így az $(\overline{SP})$ feladatot kapjuk. Ez egy $n+1$ dimenziós feladat, az alábbi szereposztásban:

$$\overline{M} = \begin{pmatrix} M & \mathbf{r} \\ -\mathbf{r}^T & 0 \end{pmatrix} \in \mathbb{R}^{(n+1)\times(n+1)}, \qquad \overline{\mathbf{q}} = \begin{pmatrix} \mathbf{0} \\ \lambda \end{pmatrix} \in \mathbb{R}^{n+1}.$$

A változók: $\overline{\mathbf{u}} = \begin{pmatrix} \mathbf{x} \\ \theta \end{pmatrix}$, $\overline{\mathbf{s}} = \begin{pmatrix} \mathbf{s} \\ \nu \end{pmatrix}$, míg a célfüggvény: $\min \lambda\theta$.

$\overline{\mathbf{u}} = \overline{\mathbf{e}}$, $\overline{\mathbf{s}} = \overline{\mathbf{e}}$ pontosan akkor belső pont, ha $\lambda = n+1$, $\mathbf{r} = \mathbf{e} - M\mathbf{e}$.

Tehát a fenti μ, λ választás esetén az $(\overline{SP})$ feladatnak az $\mathbf{e}$ vektor mindig belső pontja a fenti választás mellett. Legyen az $(\overline{SP})$ feladatnak $(\overline{\mathbf{u}}, \overline{\mathbf{s}})$ szigorúan komplementáris megoldása. Mivel az $(\overline{SP})$ feladat célfüggvénye a θ változó $(n+1)$-szerese, így ennek a változónak minden optimális megoldásban nulla az értéke, mert az optimális célfüggvényérték 0. Emiatt $(\mathbf{x}, \mathbf{s})$ – szigorúan komplementáris – megoldása az (SP) feladatnak.

2. Dikin-irány és átparaméterezés

Legyen $(\mathbf{x}, \mathbf{s})$ belső pont. Keressünk olyan $(\Delta\mathbf{x}, \Delta\mathbf{s}) \in \mathbb{R}^{2n}$ irányt, amely teljesíti az

$$M\Delta\mathbf{x} = \Delta\mathbf{s}$$

ferde altér feltételt. Az ilyen irányt az (SP) feladat *recessziós irányának* nevezzük, az $\mathbf{x}^+ := \mathbf{x} + \Delta\mathbf{x}$, $\mathbf{s}^+ := \mathbf{s} + \Delta\mathbf{s}$ pontot pedig a $(\Delta\mathbf{x}, \Delta\mathbf{s})$ recessziós irány menti *teljes lépésnek* hívjuk. Ha adott egy $\alpha > 0$ lépéshossz is, akkor az $\mathbf{x}^+ = \mathbf{x} + \alpha\Delta\mathbf{x}$, $\mathbf{s}^+ = \mathbf{s} + \alpha\Delta\mathbf{s}$ pontot *α-val tompított lépésnek* nevezzük, függetlenül attól, hogy $\alpha < 1$ teljesül, vagy sem. Világos, hogy egy $(\Delta\mathbf{x}, \Delta\mathbf{s})$ vektor pontosan akkor teljesíti a ferde altér feltételt, ha az irányában vett teljes lépés nem vezet ki az

$$M\mathbf{x} + \mathbf{q} = \mathbf{s}$$

ferde affin altérből.

Vizsgáljuk meg, hogy hogyan változik meg az $\mathbf{x}^T\mathbf{s}$ dualitásrés a lépés hatására! Felhasználva, hogy $\Delta\mathbf{x}^T\Delta\mathbf{s} = \Delta\mathbf{x}^T M\Delta\mathbf{x} = 0$, a lépés végrehajtása után a dualitásrés az

$$\left(\mathbf{x}^+\right)^T\mathbf{s}^+ = \mathbf{x}^T\mathbf{s} + \alpha\left(\mathbf{x}^T\Delta\mathbf{s} + \mathbf{s}^T\Delta\mathbf{x}\right) + \alpha^2\left(\Delta\mathbf{x}^T\Delta\mathbf{s}\right) = \mathbf{x}^T\mathbf{s} + \alpha\left(\mathbf{x}^T\Delta\mathbf{s} + \mathbf{s}^T\Delta\mathbf{x}\right)$$

alakot ölti. Ebből a dualitásrés megváltozására az

$$\left(\mathbf{x}^+\right)^T\mathbf{s}^+ - \mathbf{x}^T\mathbf{s} = \alpha\left(\mathbf{x}^T\Delta\mathbf{s} + \mathbf{s}^T\Delta\mathbf{x}\right)$$

képletet nyerjük.

Szeretnénk meghatározni egy olyan recessziós irányt, amely recessziós irányban vett teljes lépés a lehető legnagyobb dualitásrés csökkenést eredményez. Ilyen irány természetesen nem mindig létezik, hiszen az alterünk általában nem korlátos az $(\mathbf{x}, \mathbf{s})$ irány mentén. Azonban ezen a problémán könnyen segíthetünk, ha az iránykeresést megszorítjuk a ferde altér egy kompakt részhalmazára.

Ezért tekintjük – Dikin ötlete alapján – az alábbi relaxált feladatot:

$$\min\left\{\mathbf{x}\Delta\mathbf{s}+\mathbf{s}\Delta\mathbf{x} : M\Delta\mathbf{x}=\Delta\mathbf{s},\ \left\|\frac{\Delta\mathbf{x}}{\mathbf{x}}+\frac{\Delta\mathbf{s}}{\mathbf{s}}\right\|^2\le 1\right\}.$$

Ebben a relaxált iránykeresési feladatban egy szigorúan konvex, kompakt halmazon minimalizálunk egy lineáris célfüggvényt. Emiatt egyrészt egyértelműen létezik a feladat megoldása, másrészt meg fogjuk mutatni, hogy ezen – a kompakt feltételek mellett a dualitásrés csökkenését maximalizáló – irányra képlet is adható. A szóbanforgó szigorúan kompakt, konvex halmazt *Dikin-ellipszoidnak* nevezzük.

$$\mathcal{E}_D := \left\{(\Delta\mathbf{x},\Delta\mathbf{s}) : M\Delta\mathbf{x}=\Delta\mathbf{s},\ \left\|\frac{\Delta\mathbf{x}}{\mathbf{x}}+\frac{\Delta\mathbf{s}}{\mathbf{s}}\right\|^2\le 1\right\}.$$

Vezessük be a következő konstansokat, illetve konstans vektorokat:

$$\mu := \frac{\mathbf{x}^T\mathbf{s}}{n},\ \mathbf{d} := \sqrt{\frac{\mathbf{x}}{\mathbf{s}}},\ \mathbf{u} := \sqrt{\frac{\mathbf{xs}}{\mu}}.$$

Ezek a dimenziótól és a belső ponttól ugyan függenek, viszont az iránytól nem, ezért tényleg konstansnak tekinthetjük őket. Segítségükkel átparaméterezzük az iránykeresési feladatot (ahol $\mathbf{xs}$ szokásos Hadamard-szorzatot jelöli).

Az első koordináta-transzformáció:

$$\mathbf{p}_x := \sqrt{\mu}\mathbf{d}^{-1}\Delta\mathbf{x},\ \mathbf{p}_s := \sqrt{\mu}\mathbf{d}\Delta\mathbf{s} \ \Leftrightarrow\ \Delta\mathbf{x}=\frac{\mathbf{d}\mathbf{p}_x}{\sqrt{\mu}},\ \Delta\mathbf{s}=\frac{\mathbf{d}^{-1}\mathbf{p}_s}{\sqrt{\mu}}. \tag{1}$$

A második koordináta-transzformáció:

$$\mathbf{p} := \mathbf{p}_x + \mathbf{p}_s.$$

2.1. LEMMA. *A második koordináta-transzformáció inverze létezik és*

$$\mathbf{p}_x = (I+DMD)^{-1}\mathbf{p},$$
$$\mathbf{p}_s = DMD(I+DMD)^{-1}\mathbf{p},$$

ahol $D := diag(\mathbf{d})$, *azaz* d *elemeiből álló diagonális mátrix.*

Bizonyítás. Az első koordináta-transzformáció (1) miatt:

$$M\Delta\mathbf{x}=\Delta\mathbf{s} \ \Leftrightarrow\ M(\mathbf{d}\mathbf{p}_x)=\mathbf{d}^{-1}\mathbf{p}_s \ \Leftrightarrow\ \mathbf{p}_s = DMD\mathbf{p}_x,$$

tehát

$$\mathbf{p} = (I + DMD)\mathbf{p}_x.$$

Továbbá létezik $(I + DMD)^{-1}$, mert $I + DMD$-nek nincsen nulla sajátértéke, ugyanis:

$$\forall \mathbf{z} \in \mathbb{R}^n, \|\mathbf{z}\| = 1 \ : \mathbf{z}^T(I + DMD)\mathbf{z} = 1.$$

Ezzel a lemma állítását bebizonyítottuk. □

A harmadik koordináta-transzformáció:

$$\overline{\mathbf{p}} = \frac{\mathbf{p}}{\mathbf{u}} \ \Leftrightarrow \ \mathbf{p} = \overline{\mathbf{p}}\mathbf{u}. \tag{2}$$

Azért ezt az átskálázást használjuk, mert e mentén a transzformáció mentén a Dikin-ellipszoid az $\mathbb{R}^n$ egységgömbjébe, míg a lineáris célfüggvényünk az átskálázott tér lineáris függvényébe transzformálódik.

- Ferde-altér feltétel:

$$M\Delta\mathbf{x} = \Delta\mathbf{s} \ \Leftrightarrow \ DMD\mathbf{p}_x = \mathbf{p}_s \ \Leftrightarrow \ \mathbf{p} \in \mathbb{R}^n \ \Leftrightarrow \ \overline{\mathbf{p}} \in \mathbb{R}^n.$$

- Dikin-ellipszoid:

$$\left\| \frac{\Delta\mathbf{x}}{\mathbf{x}} + \frac{\Delta\mathbf{s}}{\mathbf{s}} \right\|^2 \leq 1 \ \Leftrightarrow \ \left\| \frac{\mathbf{p}_x}{\mathbf{u}} + \frac{\mathbf{p}_s}{\mathbf{u}} \right\|^2 \leq 1 \ \Leftrightarrow \ \left\| \frac{\mathbf{p}}{\mathbf{u}} \right\|^2 \leq 1 \ \Leftrightarrow \ \|\overline{\mathbf{p}}\|^2 \leq 1.$$

- Célfüggvény:

$$\mathbf{s}^T\Delta\mathbf{x} + \mathbf{x}^T\Delta\mathbf{s} = (\mu\mathbf{u})^T(\mathbf{p}_x + \mathbf{p}_s) = (\mu\mathbf{u})^T\mathbf{p} = \left(\mu\mathbf{u}^2\right)^T\overline{\mathbf{p}}.$$

Tehát a relaxált iránykeresési feladat áttranszformált alakja:

$$\min\left\{\mu\left(\mathbf{u}^2\right)^T\overline{\mathbf{p}} : \ \|\overline{\mathbf{p}}\|^2 \leq 1\right\}.$$

Vezessük be a $\mathbf{v} = \mu\mathbf{u}^2$ jelölést! A következő – szemléletesen nyilvánvaló – lemma könnyen bizonyítható.

2.2. LEMMA. *A* $\min\left\{\mathbf{v}^T\overline{\mathbf{p}} : \ \|\overline{\mathbf{p}}\|^2 \leq 1\right\}$ *feladat megoldása létezik és egyértelmű, sőt előáll az alábbi*

$$\overline{\mathbf{p}} = -\frac{\mathbf{v}}{\|\mathbf{v}\|}.$$

alakban.

Ez a mi esetünkben azt jelenti, hogy $\overline{\mathbf{p}} = -\frac{\mathbf{u}^2}{\|\mathbf{u}^2\|}$. Transzformáljuk ezt vissza az eredeti alakba.

$$\Delta\mathbf{x} = \sqrt{\mu}D(I + DMD)^{-1}\frac{-\mathbf{u}^3}{\|\mathbf{u}^2\|} =$$
$$= -D(I + DMD)^{-1}\frac{(\mathbf{xs})^{\frac{3}{2}}}{\|\mathbf{xs}\|} = -(S + XM)^{-1}\frac{(\mathbf{xs})^2}{\|\mathbf{xs}\|}.$$

Az első egyenlőséghez a koordináta-transzformációk inverzeit, a második egyenlőséghez $\mathbf{u}$ definícióját, majd a harmadikhoz az alábbi összefüggést alkalmaztuk:

$$D(I + DMD)^{-1}(XS)^{-\frac{1}{2}} = \left[(XS)^{\frac{1}{2}}D^{-1} + \left([XS]^{\frac{1}{2}}D\right)M\left(DD^{-1}\right)\right]^{-1} =$$
$$= (S + XM)^{-1},$$

ahol $X = diag(\mathbf{x})$ és $S = diag(\mathbf{s})$.

2.1. Definíció. $\Delta\mathbf{x}$ *Dikin-irány*:

$$\Delta\mathbf{x} = -(S + XM)^{-1}\frac{(\mathbf{xs})^2}{\|\mathbf{xs}\|}.$$

Dikin-iránynak nevezzük a Dikin-irány minden transzformált vektorát is.

Azaz $(\Delta\mathbf{x}, \Delta\mathbf{s}) \equiv (\mathbf{p}_x, \mathbf{p}_s) \equiv \mathbf{p} \equiv \overline{\mathbf{p}}$, ahol $\overline{\mathbf{p}} = -\frac{\mathbf{u}^2}{\|\mathbf{u}^2\|}$. Az ekvivalenciák azt jelentik, hogy a különböző koordináták ugyanannak az iránynak a koordinátái a bevezetett négy koordinátarendszerben, ezenkívül meg is kaphatók egymásból a három koordináta-transzformáció segítségével.

3. A Dikin-algoritmus és annak komplexitása

A Dikin-irányban megtett teljes Dikin-lépés dualitásrés változása:

$$\mu\left(\mathbf{u}^2\right)^T\overline{\mathbf{p}} = \mu\left(\mathbf{u}^2\right)^T\left(-\frac{\mathbf{u}^2}{\|\mathbf{u}^2\|}\right) = -\mu\|\mathbf{u}^2\| = -\|\mathbf{xs}\| < 0.$$

Az $(SP)_\varepsilon$ feladatnak konstruktív módon szeretnénk egy megoldását előállítani olymódon, hogy elindulunk egy tetszőleges $\mathbf{x}$ belső pontból, majd onnan a Dikin-irányban meglépjük az $\alpha > 0$-val módosított Dikin-lépést az $\mathbf{x}^+ = \mathbf{x} + \alpha\Delta\mathbf{x}$ pontba ($\mathbf{s}^+ = \mathbf{s} + \alpha\Delta\mathbf{s}$). Csak belső pontból lépünk, mivel a Dikin-ellipszoid nem értelmes külső- vagy határpont esetén.

Így az eddigiek alapján:

3.1. Algoritmus. Dikin-algoritmus az $(SP)_\varepsilon$ feladatra:
inicializáljuk $\mathbf{x}$-et,
amíg $\mathbf{x}^T\mathbf{s} > \varepsilon$,
számítsuk ki $\Delta\mathbf{x}$-et,
válasszuk meg α-t,
cseréljük ki $\mathbf{x}$-et $\mathbf{x} + \alpha\Delta\mathbf{x}$-re.

Az algoritmusban szereplő α lépéshossz megválasztásáról a 4. fejezetben fogunk szót ejteni.

3.1. LEMMA. *Az $0 \leq \alpha$-val korrigált Dikin-lépésre*

$$(\mathbf{x}^+)^T\mathbf{s}^+ \leq \left(1 - \frac{\alpha}{\sqrt{n}}\right)\mathbf{x}^T\mathbf{s}.$$

Bizonyítás.

$$\begin{aligned}\mathbf{x}^+ &= \mathbf{x} + \alpha\Delta\mathbf{x} = \sqrt{\mu}\mathbf{d}(\mathbf{u} + \alpha\mathbf{p}_x),\\ \mathbf{s}^+ &= \mathbf{s} + \alpha\Delta\mathbf{s} = \sqrt{\mu}\mathbf{d}^{-1}(\mathbf{u} + \alpha\mathbf{p}_x),\\ \mathbf{x}^+\mathbf{s}^+ &= \mu\left(\mathbf{u}^2 + \alpha\mathbf{u}(\mathbf{p}_x + \mathbf{p}_s) + \alpha^2\mathbf{p}_x\mathbf{p}_s\right),\end{aligned}$$

amire alkalmazva, hogy $\mathbf{p}_x + \mathbf{p}_s = \mathbf{p} = -\frac{\mathbf{u}^3}{\|\mathbf{u}^2\|}$, az α-val korrigált Dikin-lépésre:

$$\mathbf{x}^+\mathbf{s}^+ = \mu\left(\mathbf{u}^2 - \frac{\mathbf{u}^4}{\|\mathbf{u}^2\|}\alpha + (\mathbf{p}_x\mathbf{p}_s)\alpha^2\right) \tag{3}$$

egyenlet adódik. Mivel a Dikin-irány ortogonális, azaz

$$\mathbf{p}_x^T\mathbf{p}_s = 0, \text{ és } \mathbf{e}^T\frac{\mathbf{u}^4}{\|\mathbf{u}^2\|} = \|\mathbf{u}^2\|,$$

ezért az új dualitásrés:

$$\left(\mathbf{x}^+\right)^T\mathbf{s}^+ = \mu\left(\|\mathbf{u}\|^2 - \alpha\|\mathbf{u}^2\|\right) \leq \mu\left(1 - \frac{\alpha}{\sqrt{n}}\right)\|\mathbf{u}\|^2 = \left(1 - \frac{\alpha}{\sqrt{n}}\right)\mathbf{x}^T\mathbf{s}.$$

Az egyenlőtlenség a Cauchy–Bunyakovszkij–Schwartz-egyenlőtlenség miatt teljesül. Ez alapján ugyanis $\|\mathbf{u}\|^2 = \mathbf{e}^T\mathbf{u}^2 \leq \sqrt{n}\|\mathbf{u}^2\|$. □

3.1. KÖVETKEZMÉNY. *Tegyük fel, hogy $0 \leq \alpha \leq \sqrt{n}$. Ekkor a Dikin-algoritmus legfeljebb*

$$\left\lceil \log\left(\frac{\mathbf{q}^T\mathbf{x}^0}{\varepsilon}\right)\frac{\sqrt{n}}{\alpha^*}\right\rceil$$

iterációt hajt végre, ahol $\alpha^* = inf\{\alpha\}$ a választott lépéshosszak infimuma és $\mathbf{x}^0$ az algoritmus kezdőpontja.

Bizonyítás.

$$\left(1 - \frac{\alpha^*}{\sqrt{n}}\right)^k \mathbf{q}^T\mathbf{x} \leq \varepsilon,$$

$$k \geq \frac{\log\left(\frac{\varepsilon}{\mathbf{q}^T\mathbf{x}}\right)}{\log\left(1 - \frac{\alpha^*}{\sqrt{n}}\right)} \geq_{(*)} \frac{\sqrt{n}}{\alpha^*} \log\left(\frac{\mathbf{q}^T\mathbf{x}}{\varepsilon}\right),$$

ahol $(*)$ teljesül, mert

$$\log\left(1 - \frac{\alpha^*}{\sqrt{n}}\right) \leq -\frac{\alpha^*}{\sqrt{n}},$$

hiszen $1 + t \leq e^t$, ahol $t = -\frac{\alpha^*}{\sqrt{n}}$. □

4. A lépéshossz megválasztása

Túl nagy lépéshossz esetén (pl: $\alpha \approx \sqrt{n}$) előfordulhat, hogy kilépünk a belső pontok halmazából, így a következő lépésben nem tudjuk majd kiszámítani a Dikin-irányt. Az sem világos, hogy hogyan fogunk majd onnan visszatalálni a megengedettségi tartomány belsejébe.

Túl kicsi lépéshosszak esetén (azaz: $\alpha^* = 0$) az a veszély fenyeget minket, hogy nem fogjuk tudni a dualitásrést ε alá csökkenteni, esetleg még véges számú lépésben sem. Vezessük be a következő elnevezéseket!

4.1. Definíció. Az α lépéshossz *megengedett*, ha

$$\mathbf{x} + \alpha\Delta\mathbf{x},\ \mathbf{s} + \alpha\Delta\mathbf{s} > \mathbf{0},$$

és *iterációfüggetlen*, ha $\alpha = \alpha^*$ minden lépésben.

Célunk az iterációfüggetlen és megengedett α lépéshossz megkonstruálása. Ehhez szükségünk lesz a centrális úttól való eltérés mértékére, melynek segítségével értelmezni tudjuk majd a centrális út τ-környezetét. Nevezzük a konstans α lépéshosszt *τ-megfelelőnek*, ha bármely τ-környezetbeli pontból is hajtsuk végre, az α-val tompított Dikin lépés τ-környezetbeli pontot eredményez. Világos, hogy τ-megfelelő lépéshossz iteráció-független – hiszen konstans. Ezzel egyidejűleg belsőpontos is, mert a τ-környezetet a belső pontok részhalmazaként fogjuk értelmezni.

A τ-megfelelő lépéshossz egzisztenciáját fogjuk bebizonyítani.

4.2. Definíció. Legyen az $(\mathbf{x}, \mathbf{s})$ belső pont *centralitásának a mértéke*:

$$\delta(\mathbf{x}) = \delta(\mathbf{x}, \mathbf{s}) = \frac{\max(\mathbf{xs})}{\min(\mathbf{xs})}.$$

Világos, hogy $\delta(\mathbf{x}) \geq 1$, és a $\delta(\mathbf{x}) = 1$ egyenlőség azt jelenti, hogy $\mathbf{x}$ a centrális úton fekszik. Ezenkívül nyilván

$$\delta(\mathbf{x}, \mathbf{s}) = \frac{\max \mathbf{u}^2}{\min \mathbf{u}^2},$$

hiszen $\mathbf{u}^2 = \frac{\mathbf{xs}}{\mu}$.

4.3. Definíció. Legyen $\tau > 1$, $\tau \in \mathbb{R}$. A centális út *τ-környezete* azon belső pontok összessége, melyekre a centralitás mértéke nem haladja meg a τ mértéket. Másszóval

$$V_\tau = \left\{(\mathbf{x}, \mathbf{s}) \in \mathcal{F}^0 : \delta(\mathbf{x}, \mathbf{s}) \leq \tau\right\}.$$

Világos, hogy bármely $\mathbf{x}$ τ-környezetbéli ponthoz léteznek a $\tau_1, \tau_2 > 0$ pozitív valós számok úgy, hogy

$$\tau = \frac{\tau_2}{\tau_1}, \qquad \tau_2 \geq \max\left(\mathbf{u}^2\right), \qquad \tau_1 \leq \min\left(\mathbf{u}^2\right).$$

Kiemeljük, hogy ezek a számok nagyban függenek az $\mathbf{x}$ belső pont környezetbeli megválasztásától.

Emlékezzünk vissza a Dikin-algoritmus lépésbecslésében használt (3)

$$\mathbf{x}^+\mathbf{s}^+ = \mu\left(\mathbf{u}^2 - \frac{\mathbf{u}^4}{\|\mathbf{u}^2\|}\alpha + (\mathbf{p}_x\mathbf{p}_s)\alpha^2\right)$$

összefüggésre. Alkalmazzuk az alábbi jelöléseket:

$$f_\alpha(t) := t - \alpha\frac{t^2}{\|\mathbf{u}^2\|}, \qquad H := \max\{\mathbf{p}_x\mathbf{p}_s\}, \qquad h := \min\{\mathbf{p}_x\mathbf{p}_s\}.$$

Megjegyezzük, hogy értelme van vektort polinomba helyettesíteni. A környezetben maradás bizonyításához a következő formális becslést szeretnénk végrehajtani:

$$\frac{\mathbf{x}^+\mathbf{s}^+}{\mu} = \left(\mathbf{u}^+\right)^2 = \left(\underbrace{\mathbf{u}^2 - \frac{\mathbf{u}^4}{\|\mathbf{u}^2\|}\alpha}_{f_\alpha(\mathbf{u}^2)} + \underbrace{(\mathbf{p}_x\mathbf{p}_s)\alpha^2}_{h\alpha^2 \leq \sim \leq H\alpha^2}\right)$$

$$1 \underbrace{\leq}_{(C3)} \delta\left(\mathbf{x}^+, \mathbf{s}^+\right) = \frac{\max\left(\mathbf{x}^+\mathbf{s}^+\right)}{\min\left(\mathbf{x}^+\mathbf{s}^+\right)} = \frac{\max(\mathbf{u}^+)^2}{\min(\mathbf{u}^+)^2} \underbrace{\leq}_{(C1)} \frac{f_\alpha(\tau_2) + H\alpha^2}{f_\alpha(\tau_1) + h\alpha^2} \underbrace{\leq}_{(C2)} \tau.$$

Formális becslésünk érvényességéhez a $(C1)$, $(C2)$ és $(C3)$ egyenlőtlenségeket kell biztosítanunk.

- (C1) feltétele: f_α monoton növő a $[0, \tau_2]$ intervallumon.
- (C2) feltétele: $f_\alpha(\tau_2) + H\alpha^2 \leq \tau\left(f_\alpha(\tau_1) + h\alpha^2\right)$.
- (C3) feltétele: $f_\alpha(\tau_1) + h\alpha^2 > 0$.

Vegyük észre, hogy a $(C3)$ feltétel automatikusan teljesül, feltéve, hogy teljesül a $(C1)$ becslés feltétele, és ezen kívül még szigorúan teljesül a $(C2)$ becslés feltétele is. Valóban,

$$f_\alpha(\tau_1) + h\alpha^2 \underbrace{\leq}_{(C1)} f_\alpha(\tau_2) + H\alpha^2 \underbrace{\lneq}_{(C2)} \tau\left(f_\alpha(\tau_1) + h\alpha^2\right),$$

és átrendezve

$$(\tau - 1)\left(f_\alpha(\tau_1) + h\alpha^2\right) > 0.$$

Figyelembe véve, hogy $\tau > 1$, a $(C3)$ feltételt kapjuk.

Világos, hogy a $(C1)$ és a $(C3)$ feltételből következik, hogy $\mathbf{x}^+\mathbf{s}^+ > \mathbf{0}$. Ha ez minden α-nál kisebb lépéshosszra is teljesül, akkor a lépés folytonossága miatt $\mathbf{x}^+, \mathbf{s}^+ > \mathbf{0}$, tehát a tompított Dikin-lépés belső pontot eredményez. Hogy az $\mathbf{x}^+\mathbf{s}^+ > \mathbf{0}$ szorzat minden α-nál kisebb lépéshosszra is pozitív, az a $(C1)$ és $(C2)$ tulajdonságok "leszálló" jellegéből fog adódni, amit később be fogunk majd bizonyítani.

Összefoglalva, ha az α lépéshossz a τ-környezet bármelyik pontjában teljesíti a $(C1)$ és a $(C2)$ feltételeket, akkor az α lépéshossz τ-megengedett. Ez az eszmefuttatás indokolja a következő definíció bevezetését.

4.4. Definíció. Azt mondjuk, hogy az α lépéshossz kielégíti a *Dikin-feltételt a τ-környezetre vonatkozólag*, ha bármely $(\mathbf{x}, \mathbf{s})$ τ-környezetbeli pontra:

- f_α monoton növő a $[0, \tau_2]$ intervallumon, és
- $f_\alpha(\tau_2) + \alpha^2 H < \tau\left(f_\alpha(\tau_1) + \alpha^2 h\right)$ is teljesül.

Ha ez a feltétel a τ-környezetnek csak egy rögzített $(\mathbf{x}, \mathbf{s})$ pontjában teljesül, akkor azt mondjuk, hogy az α lépéshossz lokálisan teljesíti a Dikin-feltételt ebben a pontban.

Most rátérünk a Dikin-feltétel pontbeli karakterizálására.

4.1. LEMMA. *Az $\alpha > 0$ lépéshossz pontosan akkor teljesíti a lokális Dikin-feltételt, ha egyrészt $\alpha \leq \frac{\|\mathbf{u}^2\|}{2\tau_2}$, másrészt ha $\alpha < \frac{\tau - 1}{H - \tau h} \frac{\tau_1 \tau_2}{\|\mathbf{u}^2\|}$.*

Bizonyítás. f_α egy konkáv másodfokú polinom 0 és $\frac{\|\mathbf{u}^2\|}{\alpha}$ gyökökkel, tehát f_α monoton növő a $\left[0, \frac{\|\mathbf{u}^2\|}{2\alpha}\right]$ intervallumon. Így f_α pontosan akkor monoton növő a $[0, \tau_2]$-n, ha $\tau_2 \leq \frac{\|\mathbf{u}^2\|}{2\alpha}$.

A lokális Dikin-tulajdonság második feltétele:

$$\tau_2 - \alpha\frac{\tau_2^2}{\|\mathbf{u}^2\|} + \alpha^2 H = f_\alpha(\tau_2) + \alpha^2 H < \tau\left(f_\alpha(\tau_1) + \alpha^2 h\right) = \tau\left[\tau_1 - \alpha\frac{\tau_1^2}{\|\mathbf{u}^2\|} + \alpha^2 h\right].$$

Ezt az α-ban másodfokú kifejezést átrendezve

$$(H - \tau h)\alpha^2 + \frac{\tau\tau_1^2 - \tau_2^2}{\|\mathbf{u}^2\|}\alpha + \tau_2 - \tau\tau_1 < 0$$

egyenlőtlenség adódik.

Felhasználva, hogy

- $\tau_2 - \tau\tau_1 = 0$, mert $\tau = \frac{\tau_2}{\tau_1}$ és
- $\tau\tau_1^2 - \tau_2^2 = (1-\tau)\tau_1\tau_2$,

átrendezéssel a

$$(H - \tau h)\alpha < \frac{(\tau - 1)\tau_1\tau_2}{\|\mathbf{u}^2\|}$$

egyenlőtlenség adódik, ami az állítással ekvivalens, ugyanis $H - \tau h \geq 0$, mert $H \geq 0$ és $h \leq 0$, hiszen $\mathbf{p}_x^T\mathbf{p}_s = 0$. □

4.1. KÖVETKEZMÉNY. *Ha egy α lépéshossz teljesíti a lokális-, vagy globális Dikin-feltételt, akkor minden nála kisebb pozitív lépéshossz is teljesíti.*

A következő lemma bizonyítását lényegében már el is végeztük.

4.2. LEMMA. *Ha az α lépéshosszra teljesül a lokális/globális Dikin-feltétel, akkor*

i. $\mathbf{x}^+, \mathbf{s}^+ > \mathbf{0}$,

ii. $\delta(\mathbf{x}^+) \leq \tau$.

Tehát az α lépéshossz lokálisan/globálisan τ-megfelelő.

Bizonyítás. A Dikin-feltétel teljesüléséből adódik, hogy

$$\delta\left(\mathbf{x}^+\right) = \frac{\max\left(\mathbf{x}^+\mathbf{s}^+\right)}{\min\left(\mathbf{x}^+\mathbf{s}^+\right)} \leq \frac{\mu\left(f_\alpha(\tau_2) + \alpha^2 H\right)}{\mu\left(f_\alpha(\tau_1) + \alpha^2 h\right)} \leq \tau,$$

azaz a lemma második felét igazoltuk. Továbbá (3)-ból, f_α és h definíciójából következik, hogy

$$\mathbf{x}^+\mathbf{s}^+ \geq \left(f_\alpha(\tau_1) + \alpha^2 h\right)\mathbf{e}.$$

Azt kell megmutatni, hogy ez az alsó korlát pozitív. A Dikin-lépés folytonossága miatt ebből már következik, hogy $\mathbf{x}^+ \in \mathcal{F}^0$. A Dikin-feltétel felhasználásval:

$$\tau\left(f_\alpha(\tau_1) + \alpha^2 h\right) > f_\alpha(\tau_2) + \alpha^2 H > f_\alpha(\tau_1) + \alpha^2 h.$$

Átrendezve kapjuk, hogy

$$(\tau - 1)\left(f_\alpha(\tau_1) + \alpha^2 h\right) > 0.$$

Miután $\tau > 1$, így az állítást beláttuk. □

Hátramaradt még a globálisan τ-megfelelő lépéshossz létezésének a bizonyítása.

4.3. LEMMA. *Legyen* $\mathbf{x}$ *a* τ*-környezet tetszőleges pontja. Ekkor:*

i. $\frac{1}{\tau\sqrt{n}} \leq \frac{\|\mathbf{u}^2\|}{2\tau_2}$,

ii. $\frac{4(\tau-1)}{\tau+1}\frac{1}{\tau\sqrt{n}} < \frac{\tau-1}{H-\tau h}\frac{\tau_1\tau_2}{\|\mathbf{u}^2\|}$.

Bizonyítás.

i.

$$\frac{\|\mathbf{u}^2\|}{2\tau_2} \geq \frac{\|\tau_1\mathbf{e}\|}{2\tau_2} = \frac{\tau_1\sqrt{n}}{2\tau_2} \geq \frac{\tau_1}{\tau_2\sqrt{n}} = \frac{1}{\tau\sqrt{n}}.$$

ii.

- $H - \tau h \leq (\tau + 1)\|\mathbf{p}_x\mathbf{p}_s\|_\infty$, ami H és h definíciójából következik.
- $\|\mathbf{p}_x\mathbf{p}_s\|_\infty \leq \frac{\|\mathbf{p}\|^2}{4}$, ugyanis $\mathbf{p}_x\mathbf{p}_s = \frac{\left((\mathbf{p}_x+\mathbf{p}_s)^2-(\mathbf{p}_x-\mathbf{p}_s)^2\right)}{4}$, így a következő becslés nyerhető:

$$-\frac{(\mathbf{p}_x - \mathbf{p}_s)^2}{4} \leq \mathbf{p}_x\mathbf{p}_s \leq \frac{(\mathbf{p}_x + \mathbf{p}_s)^2}{4}.$$

  Miután $\mathbf{p}_x^T\mathbf{p}_s = 0$, ezért $\|\mathbf{p}_x + \mathbf{p}_s\|^2 = \|\mathbf{p}_x - \mathbf{p}_s\|^2$. Ebből következik, hogy

$$\mathbf{p}_x\mathbf{p}_s \leq \frac{\|\mathbf{p}_x + \mathbf{p}_s\|^2}{4}\mathbf{e}.$$

- $\|\mathbf{p}\|^2 \leq \tau_2$, mivel

$$\|\mathbf{p}\|^2 = \left\|\frac{\mathbf{u}^3}{\|\mathbf{u}^2\|}\right\| \leq \|\mathbf{u}\|_\infty^2 \left\|\frac{\mathbf{u}^2}{\|\mathbf{u}^2\|}\right\|^2 \leq \tau_2.$$

- $\left\|\mathbf{u}^2\right\| \leq \left\|\mathbf{u}^2\right\|_\infty \|\mathbf{e}\| \leq \tau_2\sqrt{n}$.

Így

$$\frac{\tau - 1}{H - \tau h}\frac{\tau_1\tau_2}{\|\mathbf{u}^2\|} \geq \frac{4(\tau-1)}{(\tau+1)\tau_2}\frac{\tau_1\tau_2}{\sqrt{n}\tau_2} = \frac{4(\tau-1)}{\tau+1}\frac{1}{\tau\sqrt{n}}.$$

Ezzel bebizonyítottuk a lemmát. □

4.2. KÖVETKEZMÉNY. *Minden $\tau > 1$ számra:*

$$0 < \min\left\{\frac{1}{\tau\sqrt{n}}, \frac{4(\tau-1)}{\tau+1}\frac{1}{\tau\sqrt{n}}\right\} \le \min\left\{\frac{\|\mathbf{u}^2\|}{2\tau_2}, \frac{\tau-1}{H-\tau h}\frac{\tau_1\tau_2}{\|\mathbf{u}^2\|} \;:\; \mathbf{x}\in V_\tau\right\}.$$

Ezen kívül bármely $\alpha > 0$ lépéshossz globálisan τ-megfelelő, ha

$$0 < \alpha \le \min\left\{\frac{1}{\tau\sqrt{n}}, \frac{4(\tau-1)}{\tau+1}\frac{1}{\tau\sqrt{n}}\right\}.$$

Megjegyzés.

$$\frac{1}{\tau\sqrt{n}} = \frac{4(\tau-1)}{\tau+1}\frac{1}{\tau\sqrt{n}} \iff \tau = \frac{5}{3}.$$

4.1. TÉTEL. *Legyen $\mathbf{x}^0$ belső pont. Ha $\mathbf{x}^0$ a centrális út valamelyik pontja, akkor válasszuk a $\tau > 1$ környezetet tetszőlegesen. Ha $\mathbf{x}^0$ nem a centrális úton van, akkor legyen $\tau = \delta(\mathbf{x})$.*

$$\alpha := \begin{cases} \frac{4(\tau-1)}{\tau+1}\frac{1}{\tau\sqrt{n}} & 1 < \tau \le \frac{5}{3}, \\ \frac{1}{\tau\sqrt{n}} & \tau > \frac{5}{3}. \end{cases}$$

Ekkor az $\mathbf{x}^0$ pontból indított Dikin-algoritmus az $(SP)_\varepsilon$ feladatot legfeljebb

$$\begin{cases} \left\lceil \frac{(\tau+1)}{4(\tau-1)} n\tau \log\left(\frac{\mathbf{q}^T\mathbf{x}^0}{\varepsilon}\right)\right\rceil & 1 < \tau \le \frac{5}{3}, \\ \left\lceil n\tau \log\left(\frac{\mathbf{q}^T\mathbf{x}^0}{\varepsilon}\right)\right\rceil & \tau > \frac{5}{3}. \end{cases}$$

iterációban megoldja.

Bizonyítás. A 4.2. következmény alapján α globálisan τ-megfelelő lépéshossz. Valóban, hiszen

$$\frac{1}{\tau\sqrt{n}} < \frac{4(\tau-1)}{\tau+1}\frac{1}{\tau\sqrt{n}}$$

pontosan akkor teljesül, ha $\tau > \frac{5}{3}$. Ezért α belsőpontos is, azaz belső pontból belső pontba jutunk, így a tompított Dikin-lépés minden iterációban belsőpontos megoldást generál, csökkentett dualitásréssel. Megálláskor a dualitásrés nem haladja meg az ε hibakorlátot, tehát a Dikin-algoritmus az $(SP)_\varepsilon$ feladat megoldását szolgáltatja.

A lépésszám a kezdeti dualitásrés és az ε hibakorlát arányának a bitleírásában polinomiális.

Valóban, az 3.1. következmény, α iterációfüggetlensége és definíciója a kívánt

$$\left\lceil \log\left(\frac{\mathbf{q}^T\mathbf{x}^0}{\varepsilon}\right)\frac{\sqrt{n}}{\alpha^*}\right\rceil = \left\lceil \log\left(\frac{\mathbf{q}^T\mathbf{x}^0}{\varepsilon}\right)\frac{\sqrt{n}}{\alpha}\right\rceil = \begin{cases} \left\lceil \frac{(\tau+1)}{4(\tau-1)} n\tau \log\left(\frac{\mathbf{q}^T\mathbf{x}^0}{\varepsilon}\right)\right\rceil & 1 < \tau \le \frac{5}{3}, \\ \left\lceil n\tau \log\left(\frac{\mathbf{q}^T\mathbf{x}^0}{\varepsilon}\right)\right\rceil & \tau > \frac{5}{3} \end{cases}$$

lépésszámkorlátot szolgáltatja. □

5. Összefoglalás

Az [1] könyv 456. oldalán található Lemma E.4 bizonyításában az (E.16) lépés hibás, mert az összehasonlított vektorok minimális és maximális értéke nem feltétlenül ugyanazon koordinátán vétetnek fel. Ezt a lépést kijavítva egy élesebb tételhez jutottunk, miközben az eredeti bizonyítás gondolatmenete továbbra is alkalmazható volt.

Fontos megjegyezni, hogy a javítással nem változott meg az algoritmus komplexitása, illetve az eredeti tétel állítása érvényben maradt.

Hivatkozások

[1] Cornelis Roos, Tamás Terlaky, Jean-Philippe Vial: *Interior Point Methods for Linear Optimization*, Second Edition, 65–70, 451–459; Springer, 2005

(Beérkezett: 2008. június 17.)

MIKLÓS ZOLTÁN
Eötvös Loránd Tudományegyetem, Természettudományi Kar, Operációkutatási Tanszék
1117 Budapest, Pázmány Péter Sétány 1/c.
miklosz@cs.elte.hu

TAKÁCS SZABOLCS
Eötvös Loránd Tudományegyetem, Természettudományi Kar, Operációkutatási Tanszék
1117 Budapest, Pázmány Péter Sétány 1/c.
tretarkhon@gmail.com

IMPROVEMENT ON THE PROOF OF 'A POLYNOMIAL DIKIN-TYPE PRIMAL-DUAL ALGORITHM FOR LINEAR PROGRAMMING'

Zoltán Miklós and Szabolcs Takács

The Dikin ellipsoid method for linear programming is presented in our paper on the basis of [1]. We correct a mistake in the proof of a technical lemma for the complexity analysis of the algorithm, then we propose a new selection rule for the environment parameter by correcting the mistake in question. Further, we determine an iteration independent steplength and prove an upper bound for the complexity.

Alkalmazott Matematikai Lapok **26** (2009), 381-402.

MONOTON KÉMIAI REAKCIÓHÁLÓZATOK

PATRICK DE LEENHEER, DAVID ANGELI, EDUARDO D. SONTAG
Fordította: Várdai Judit[1]

Elemzünk néhány kémiai rekcióhálózatot, és megmutatjuk, hogy minden megoldás konvergál valamely egyensúlyi helyzethez. A kinetikáról feltételezhető, hogy monoton, de egyébként tetszőleges. Ha a diffúzió hatását figyelembe vesszük, a következtetések változatlanok maradnak. Vizsgálatunk legfontosabb eszközei a monoton dinamikai rendszerek elméletéből származnak. Ennek az elméletnek néhány jellegzetességét fogjuk áttekinteni, és egy olyan speciális vonzással kapcsolatos eredmény önálló bizonyítását adjuk, amit fő eredményünk bizonyításához is használunk.

1. Bevezetés

A kémiai reakciók dinamikai viselkedésének elméleti tanulmányozásában a kutatás egy eredményes és nagyon hatékony időszaka az utolsó néhány évtized. Ennek a folytonos figyelemnek egy sajátos oka lehet az, hogy jelenleg nem áll rendelkezésre olyan egységes elmélet, amely tetszőleges topológiájú reakcióhálózatokra és tetszőleges kinetikára vonatkozna. De ha vagy a topológiában, vagy a kinetikában megszorításokat teszünk, akkor meglehetősen általános eredményeket lehet kapni. Például az a jelentős munka, ami ma Feinberg–Horn–Jackson-elméletnek ismert, [11, 19] – és [7, 29] a legújabb eredmény, – a reakciósebességet korlátozza a kinetikai tömeghatásra, de egészen általános topológiát is figyelembe vesz. A tömeghatás típusú kinetika feltételezése lehetővé teszi Ljapunov-függvény konstruálását, ami által levonhatjuk azt a következtetést, hogy a megoldások konvergálnak. Másrészről korlátozható a hálózat topológiája, de feloldható a tömeghatás típusú kinetika feltételezése. Egy ilyen enyhítés feltételezi, hogy ezek monoton függvények (a reakció reagenseinek koncentrációjában). Például [21]-ben a következő hálózatot tanulmányozták:

$$A_1 + A_2 \rightleftharpoons B_1 \rightleftharpoons B_2 \rightleftharpoons \dots B_n,$$

[1] A dolgozat a következő fordítása: P. De Leenheer, D. Agneli, E. D. Sontag, "Monotone chemical reaction networks", Journal of Mathematical Chemistry, 41(3)(2007) 295–314.

amiben csak az első reakciólépés bimolekuláris, és a többi monomolekuláris. Abban a cikkben megmutatták, hogy ennek a hálózatnak a megoldásai egy bizonyos monotonitási tulajdonsággal rendelkeznek, olyan értelemben, amit később megmagyarázunk. Ezzel rokon ötletet találunk [30]-ban, ahol elmagyarázzák, hogy bizonyos reakcióhálózatokat hogyan kell transzformálni úgynevezett kooperatív rendszerekké (ezeket is tárgyaljuk később).

Ebben a cikkben célunk, hogy megmutassuk, hogy hogyan lehet globális konvergenciára vonatkozó eredményeket kapni egy speciális topológiájú hálózatra – amely tartalmazza és általánosítja a fenti hálózatot arra az esetre, amikor a kinetika monoton, de egyébként tetszőleges. Továbbá, az eredményeink érvényesek maradnak, ha figyelembe vesszük a diffúzió hatásait is, ily módon [23] eredményeit általánosítjuk, amik az $A_1 + A_2 \rightleftharpoons B$ megfordítható reakcióra vonatkoznak.

Hogy tisztán lássuk, a monotonitás miért játszhat szerepet a kémiai reakciókkal kapcsolatban, át fogjuk tekinteni a monoton rendszerek fogalmát, és kiemelünk néhányat ezek tulajdonságaiból. Két évtizeddel ezelőtt M. W. Hirsch kezdte el kiépíteni a monoton rendszerek elméletét a [12, 13, 14, 15, 16] cikksorozatban; de lásd még H. L. Smith kiváló összefoglalóját [27] is. Általánosságban a monoton dinamikai rendszer egy Φ folytonos félfolyam az X metrikus téren, amely úgy van ellátva egy $\preceq$ kompatibilis parciális rendezéssel, hogy a folyam megőrzi a parciális rendezést:

$$\forall x, y \in X; \quad x \preceq y \Rightarrow \Phi_t(x) \preceq \Phi_t(y), \ \forall t \in \mathbb{R}_+. \tag{1}$$

Tekintsük a következő differenciálegyenlet-rendszert:

$$\dot{x}(t) = f(x(t)) \quad (x(t) \in \mathbb{R}^n),$$

ahol $f \in \mathcal{C}^1$ olyan vektormező, amelyikről feltételezzük, hogy előre nézve teljes. Azaz a teljes megoldás értelmezési tartományának szuprémuma $+\infty$. (Habár az alábbiak akár az állapottérre, akár a vektormező simaságára vonatkozó sokkal gyengébb kikötések mellett is érvényesek.)

Azonnal felmerül a kérdés, hogy ez mikor generál monoton dinamikai rendszert valamilyen nem triviális rendezésre nézve. Erre a kérdésre nehéz megtalálni a választ. Habár amikor adott egy parciális rendezés, és azt kérdezzük, hogy a rendszer monoton-e valamely $K \subset \mathbb{R}^n$ kúp által generált parciális rendezésre nézve, ilyen esetekben van módszer a monotonitás ellenőrzésére. (Azt mondjuk, hogy $K \subset \mathbb{R}^n$ kúp, ha K olyan nem üres, zárt halmaz, amelyre $K + K \subset K$, $\mathbb{R}_+ K \subset K$ és $K \cap (-K) = \{0\}$ teljesül.) A következőkben áttekintünk néhány ilyen tesztet.

A legismertebb példa valószínűleg az, amikor f *kooperatív*, ami azt jelenti, hogy az f' Jacobi-mátrix főátlóján kívüli elemek nemnegatívak. Jól ismert, hogy ebben az esetben az $\dot{x}(t) = f(x(t))$ rendszer által generált folyam monoton, mivel megőrzi a szokásos komponensenkénti rendezést $\mathbb{R}^n$-ben, lásd például a 3.1.1. tételt és a 3.1.1. megjegyzést [27]-ben. Precízebben: ezt a rendezést az $\mathbb{R}^n_+$ ortáns kúp generálja $\mathbb{R}^n$-ben:

$$x \preceq y \Leftrightarrow y - x \in \mathbb{R}^n_+.$$

Ez általánosítható azokra az esetekre, amikor a parciális rendezést $\mathbb{R}^n$ bármelyik $\mathcal{O}$ ortáns kúpja generálja. Ilyenkor a rendezést a következőképpen definiáljuk:

$$x \preceq_{\mathcal{O}} y \Leftrightarrow y - x \in \mathcal{O}. \tag{2}$$

A monotonitás ellenőrzésére ebben az esetben egy egyszerű grafikai ellenőrzés áll rendelkezésre, lásd [27, 49. oldal]. Ez annak ellenőrzését jelenti, hogy a rendszer incidenciagráfja nem tartalmaz-e negatív paritású hurkot. (A rendszer incidenciagráfja n pontból áll, mindegyik az állapotvektor egy komponensét reprezentálja, és előjeles élek kötik össze a pontokat: a j-edik csomópontból az i-edikbe mutató élhez a $\partial_j f_i$ parciális derivált függvény előjelét rendeljük hozzá. Ez természetesen megköveteli, hogy a derivált ne változtasson előjelet, és legalább egy pontban különbözzék nullától. Egy hurok paritása egyszerűen a hurkot alkotó éleken lévő előjelek szorzata; ennél a tesztnél a hurokéleket figyelmen kívül hagyjuk.)

Ha a parciális rendezést egy tetszőleges $K \subset \mathbb{R}^n$ kúp generálja ((2)-ben egyszerűen helyettesítsük $\mathcal{O}$-t K-val), akkor is ellenőrizhető a monotonitás, habár a teszt többé már nem grafikus [17, 30, 2].

A legfontosabb ok, amiért a monoton rendszereket olyan kiterjedten tanulmányozzák, valószínűleg az, hogy sokat tudhatunk meg az aszimptotikus viselkedésükről. Közelítőleg azt mondhatjuk, hogy a legtöbb megoldás konvergál az egyensúlyok halmazához. De ebben az összefüggésben két kérdést érdemes megemlíteni. Először is, a legtöbb meglévő konvergenciakritérium erősebb monotonitási feltételt követel meg, mint (1). Jellemzően feltételezik, hogy a félfolyam *erősen rendezéstartó*, lásd [27, 2. oldal], vagy (esetlegesen) *erősen monoton* – amiből következik a korábbi – lásd ugyanazon hivatkozás 3. oldalát a pontos definíciókért. Ennek a feltételnek az ellenőrzése a gyakorlatban gyakran nem túl könnyű, vagy ami még rosszabb: lehet, hogy a rendszer monoton, de mégsem tesz eleget ezeknek az erősebb feltételeknek. Másodszor, ezeknek az eredményeknek a bizonyítása nem triviális, és alapvető eszközöket igényel a monoton rendszerek elméletéből.

Ezekhez képest kivételes partikuláris esetet tudtunk kezelni [20]-ban, ahol egy az $\mathbb{R}^n$-ben kooperatív, egyetlen egyensúllyal bíró rendszer globális aszimptotikus stabilitását vizsgáltuk. Annak a bizonyításnak a gondolatmenetét általánosítjuk a B Függelékben egyetlen egyensúllyal rendelkező monoton folytonos félfolyamokra. Ez az eredmény hasznos lehet végtelen dimenziós rendszerekre is (amilyenek a késleltetett egyenletek). Azonkívül az itt adott bizonyítás önmagában teljes.

Egy példán megmutatjuk, hogy egy partikuláris kémiai reakció minden megoldása konvergál egy egyensúlyhoz. A monoton rendszerek elméletének alkalmazásaira további példák találhatóak a kemosztát modellekre vonatkozó irodalomban [28]. Például egy változó hozamú modell monoton rendszerré transzformálható úgy, hogy a rendezés nem a szokásos komponensenkénti rendezés $\mathbb{R}^n$-en. [10]-ben egy hasonló, de többféle táplálékot tartalmazó modell analíziséhez is kiaknázzuk ezt a transzformációt.

A monoton dinamikai rendszereket újabban kiterjesztettük monoton I/O rendszerekké [2]-ben, hogy megkönnyítsük az ilyen részrendszerekből álló rendszerek

tanulmányozását (kaszkádok, visszacsatolás). Utalunk [4, 3, 5, 6, 8, 9]-re, amikben ennek az elméletnek továbbfejlesztése és alkalmazásai találhatók, a molekuláris biológia, az ökológia és a kémiai reakcióhálózatok területéről vett példákkal.

2. Egy kémiai reakció

Tekintsük a következő reakciót:

$$\mathcal{C}_1 \rightleftharpoons \cdots \rightleftharpoons \mathcal{C}_{i-1} \rightleftharpoons \mathcal{C}_i \rightleftharpoons \mathcal{C}_{i+1} \rightleftharpoons \cdots \rightleftharpoons \mathcal{C}_{n+1},$$

ahol minden $\mathcal{C}_i$ komplex különböző kémiai anyagfajták súlyozott összege a következőként:

$$\mathcal{C}_i = \sum_{k=1}^{n_i} a_i^k X_i^k$$

valamilyen pozitív a_i^k egészekre.

Ezeknek a hálózatoknak néhány speciális esetét tanulmányoztuk [5]-ben (ahol minden komplex pontosan egy anyagfajtát tartalmaz, és a hálózatban minden komplex különböző) és [21]-ben (ahol $\mathcal{C}_1 = X_1 + X_2$ két anyagfajtából, és minden rákövetkező komplex pontosan egy anyagfajtából áll, minden komplex a hálózatban különböző, és a kinetikus tömeghatás törvényét feltételezik).

Ebben a cikkben feltételezzük, hogy legalább egy komplex nem triviális, azaz létezik legalább egy $n_i > 1$. Azt is feltesszük, hogy minden anyagfajta pontosan egy komplex része, vagyis $X_i^k \neq X_j^l$ minden k, l-re, ha $i \neq j$. A $\mathcal{C}_i$ komplexhez társított koncentrációvektort $x_i = (x_i^1, \ldots, x_i^{n_i})^\top$ jelöli, a hozzá társított sztöchiometriai vektort pedig $a_i = (a_i^1, \ldots, a_i^{n_i})^\top$. Alkalmazni fogjuk még a teljes koncentrációvektort, ami $x = (x_1^\top, \ldots, x_{n+1}^\top)^\top$, ahol $x \in \mathbb{R}_+^N$, és N az összes n_i összege: $N := \sum_{i=1}^{n+1} n_i$.

Az összes reakciósebességről feltételezzük, hogy monoton, folytonosan differenciálható függvénye a reaktáns anyagfajták koncentrációjának. Ez 0, ha egy anyagfajta hiányzik, és pozitív, ha az összes reaktáns anyagfajta jelen van. A $\mathcal{C}_i \rightleftharpoons \mathcal{C}_{i+1}$ reakciólépés előremutató sebességét R_i jelöli, a hátramutató sebesség R_{-i}. Formálisan, minden $i = 1, \ldots, n$-re a cikk további részében feltételezzük, hogy:

1. $R_i : \mathbb{R}_+^{n_i} \to \mathbb{R}_+$,
2. $\forall x_i \in \partial\mathbb{R}_+^{n_i}, \quad R_i(x_i) = 0$,
3. $\forall x_i \in \operatorname{int}(\mathbb{R}_+^{n_i}) \quad R_i(x_i) > 0$ és $(R_i'(x_i))^T \in \operatorname{int}(\mathbb{R}_+^{n_i})$,

és hasonlóan az R_{-i} visszafelé haladó reakció sebességére. (Vegyük észre, hogy az R_{-i} sebesség $x_{i+1} \in \partial\mathbb{R}_+^{n_{i+1}}$ esetén van definiálva.)

Ismert példa a *kinetikus tömeghatás törvénye*, ahol a reakció sebessége $R_i(x_i) = k_i \Pi_{k=1}^{n_i} (x_i^k)^{a_i^k}$ valamilyen $k_i > 0$ mellett kielégíti a feltételeket.

Definiáljuk a reakció sebességének vektorát a következőképp:

$$R(x) = (R_1(x_1), R_{-1}(x_2), \ldots R_n(x_n), R_{-n}(x_{n+1}))^\top$$

és a hálózat sztöchiometriai mátrixát:

$$S = \begin{pmatrix} -a_1 & +a_1 & 0 & 0 & \dots & 0 \\ +a_2 & -a_2 & -a_2 & +a_2 & & \\ \vdots & & & \ddots & & \dots \\ 0 & \dots & +a_n & -a_n & -a_n & +a_n \\ 0 & \dots & 0 & 0 & +a_{n+1} & -a_{n+1} \end{pmatrix}.$$

Ezzel a koncentrációkra vonatkozó differenciálegyenlet:

$$\dot{x}(t) = SR(x(t)). \tag{3}$$

A szokásos érvelés mutatja, hogy a (3) rendszer pozitív, azaz az $\mathbb{R}^N_+$ nemnegatív ortáns pozitív invariáns halmaz. Megjegyezzük, hogy ez a rendszer nem monoton $\mathbb{R}^N$ egyetlen ortánsa által generált rendezésre sem. Ez a (3) rendszer incidenciagráfjának vizsgálatából látható, amely tartalmaz negatív paritású hurkot. Valóban, tekintsünk egy olyan hurkot, amelyiket egyrészt két olyan csomó alkot, amelyek azonos komplexben található anyagfajtáknak felelnek meg, és másrészt egy harmadik, amelyik a szomszédos komplexben (ez egy olyan komplex, amelyik az elsőből egyetlen reakciólépéssel elérhető) található anyagfajtának felel meg. Világos, hogy az ilyen csomó negatív paritású. Fő eredményünk a következő:

1. TÉTEL. *A (3) rendszer minden megoldása egyensúlyi ponthoz konvergál.*

A következő vizsgálatunkban feltételezzük, hogy van legalább egy olyan komplex, amelynek összes alkotó anyagfajtája nullától különböző kezdeti koncentrációval van jelen:

$$\exists i : x_i^k(0) > 0, \qquad \forall k = 1, \dots, n_i. \tag{4}$$

Ha ugyanis (4) nem állna fenn, akkor egyetlen egy reakció sem menne végbe. Megjegyezzük, hogy az ilyen kezdeti koncentrációk olyan egyensúlynak felelnek meg, amelyekre a 1. tétel triviálisan teljesül, így az általánosság megszorítása nélkül feltételezhetjük (4)-et.

Minden olyan C_i komplexhez, amelyre $n_i > 1$, létezik $n_i - 1$ független lineáris első integrál.

Valóban:

$$\frac{d}{dt}\left(\frac{x_i^k}{a_i^k} - \frac{x_i^1}{a_i^1}\right) = 0 \qquad \forall k = 2, \dots, n_i \tag{5}$$

(3) megoldásai mentén, és így kapjuk:

$$x_i^k(t) = \beta_i^k x_i^1(t) + \alpha_i^k \qquad \forall k = 2, \dots, n_i \tag{6}$$

valamilyen $\alpha_i^k \in \mathbb{R}$ (ami függ a kezdeti feltételektől) és $\beta_i^k := \frac{a_i^k}{a_i^1} > 0$ számokkal. Valóban, az általánosság megszorítása nélkül feltételezhetjük, hogy:

$$\alpha_i^k \geq 0, \qquad \forall k = 2, \dots, n_i.$$

Ahhoz, hogy ezt belássuk, vegyük észre, hogy – esetleg az anyagfajtáknak az egyes komplexeken belüli átcímkézése után – fennáll, hogy:

$$\frac{x_i^k(0)}{a_i^k} \geq \frac{x_i^1(0)}{a_i^1}, \qquad \forall k = 2, \ldots, n_i,$$

amiből állításunk közvetlenül következik.

(6) miatt elegendő minden C_i komplexben az x_i^1 *első* anyagfajta koncentrációjának dinamikáját tekinteni. Minden i-re definiáljuk:

$$\begin{aligned} y_i &:= x_i^1, \\ r_i(y_i) &:= R_i(y_i, \beta_i^2 y_i + \alpha_i^2, \ldots, \beta_i^{n_i} y_i + \alpha_i^{n_i}), \\ r_{-i}(y_{i+1}) &:= R_{-i}(y_{i+1}, \beta_{i+1}^2 y_{i+1} + \alpha_{i+1}^2, \ldots). \end{aligned}$$

Megjegyezzük, hogy minden r_i függvény folytonosan differenciálható a következő tulajdonságokkal:

$r_1 : \mathbb{R}_+ \longrightarrow \mathbb{R}_+, \quad r_i(0) = 0, \quad r_i(y_i) > 0 \quad$ és $\quad r_i'(y_i) > 0 \quad \forall y_i > 0$, és hasonlóan minden r_{-i} -re is.

Legyen $y := (y_1, \ldots, y_{n+1})^T, r(y) := (r_1(y_1), r_{-1}(y_2), \ldots, r_n(y_n), r_{-n}(y_{n+1}))^T$, és legyen:

$$\tilde{S} = \begin{pmatrix} -a_1^1 & +a_1^1 & 0 & 0 & \ldots & 0 \\ +a_2^1 & -a_2^1 & -a_2^1 & +a_2^1 & \ldots & 0 \\ \vdots & & & \ddots & & \ldots \\ 0 & \ldots & +a_n^1 & -a_n^1 & -a_n^1 & +a_n^1 \\ 0 & \ldots & 0 & 0 & +a_{n+1}^1 & -a_{n+1}^1 \end{pmatrix},$$

így a következő rendszerhez jutunk:

$$\dot{y}(t) = \tilde{S} r(y(t)),$$

ahol $y \in \mathbb{R}_+^{n+1} \setminus \{0\}$, (megjegyezve, hogy (4) miatt a 0-t kizárjuk).

Mivel $y_1(t)/a_1^1 + y_2(t)/a_2^1 + \cdots + y_{n+1}(t)/a_{n+1}^1 = C$ valamilyen $C > 0$ -ra a megoldások mentén, a dimenziót 1-gyel tudjuk csökkenteni, ha elhagyjuk az y_{n+1} egyenletét, és ezután n új változót vezetünk be:

$$z_j = \sum_{i=1}^{j} \frac{y_i}{a_i^1}, \qquad j = 1, \ldots, n.$$

Az inverz transzformáció:

$$\begin{aligned} y_1 &= a_1^1 z_1 \\ y_j &= a_j^1 (z_j - z_{j-1}), \quad j = 2, \ldots, n. \end{aligned}$$

Ezeket az új koordinátákat alkalmazva kapjuk a redukált rendszer egyenleteit:

$$\begin{aligned}
\dot{z_1} &= -r_1(a_1^1 z_1) + r_{-1}(a_2^1(z_2 - z_1)) \\
&\vdots \\
\dot{z_k} &= -r_k(a_k^1(z_k - z_{k-1})) + r_{-k}(a_{k+1}^1(z_{k+1} - z_k)) \quad k = 2, \ldots, n-1 \qquad (7) \\
&\vdots \\
\dot{z_n} &= -r_n(a_n^1(z_n - z_{n-1})) + r_{-n}(a_{n+1}^1(C - z_n))
\end{aligned}$$

a kompakt, konvex

$$\Omega := \{z \in \mathbb{R}^n | 0 \leq z_1 \leq z_2 \leq \cdots \leq z_n \leq C\}$$

állapottérrel. Nyilvánvaló, hogy a (7) rendszer kooperatív (és tridiagonális). Azaz $\partial_j g_k(z) > 0$, ha g jelöli (7) jobb oldalát.

1. LEMMA. *Ha $z^* \in \Omega$ a (7) rendszernek egyensúlyi helyzete, akkor $z^* \in \mathrm{int}(\Omega)$. Továbbá, z^* hiperbolikus és lokálisan aszimptotikusan stabilis egyensúlyi helyzet.*

Bizonyítás. Tegyük fel, hogy $z^* \in \partial\Omega$ a (7) rendszer egyensúlyi helyzete. Ekkor vagy $z_1^* = 0$, vagy $z_n^* = C$, vagy $z_k^* = z_{k+1}^*$ valamely $k \in \{1, ..., n-1\}$ mellett. Felhasználva, hogy minden r_i és r_{-i} függvény csak 0-ban lehet 0, minden egyes esetben ellentmondásra jutunk azzal, hogy $C > 0$. Ezzel bebizonyítottuk az első részt.

A második részhez megjegyezzük, hogy a Jacobi-mátrix egyensúlyi helyzetben a következő szerkezetű:

$$J = \begin{pmatrix}
-a_1^1 - a_2^1 & +a_2^1 & 0 & \ldots & 0 \\
+a_1^2 & -a_1^2 - a_3^2 & +a_3^2 & \ldots & 0 \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
0 & \ldots & +a_{(n-2)}^{(n-1)} & -a_{(n-2)}^{(n-1)} - a_n^{(n-1)} & +a_n^{(n-1)} \\
0 & \ldots & 0 & +a_{(n-1)}^n & -a_{(n-1)}^n - a_n^n
\end{pmatrix},$$

ahol minden $a_j^i > 0$.

J diagonálisan dominált, és az A függelékben majd bebizonyítjuk, hogy ebből következik, hogy Hurwitz-típusú mátrix.

Emlékezzünk arra, hogy a B $n \times n$ -es mátrixot diagonálisan dominánsnak nevezzük, ha létezik n olyan $d_i > 0$ szám, hogy

$$b_{ii}d_i + \sum_{j \neq i} |b_{ij}|d_j < 0, \quad \forall i = 1, \ldots, n.$$

Egy kooperatív mátrixnál, mint amilyen J, a fenti definícióból elhagyható az abszolút érték. Következésképpen találnunk kell egy d vektort pozitív koordinátákkal, hogy a Jd vektor minden koordinátája negatív legyen. Vegyük észre, hogy $J\mathbf{1}$ – ahol $\mathbf{1}$ olyan vektor, amelynek minden koordinátája 1 – első és utolsó koordinátája negatív ($-a_{11}$, illetve $-a_{nn}$) és az összes többi 0. Ez azt sugallja, hogy d-t, mint az $\mathbf{1}$ vektor megfelelő perturbációját keressük.

A következő rekurzív formulával definiáljuk az $(n-1)$ számú ε_j paramétert:

$$0 < \varepsilon_1 < \frac{a_{11}}{a_{11} + a_{12}},$$
$$0 < \varepsilon_j < \varepsilon_{j-1} \frac{a_{j(j-1)}}{a_{j(j-1)} + a_{j(j+1)}}, \quad j = 2, \ldots, n-1.$$

Nyilvánvaló, hogy $\varepsilon_j < 1$ minden $j = 1, \ldots, n-1$ esetén. Legyen a d vektor definíciója a következő:

$$d_i := 1 - \varepsilon_i, \quad i = 1, \ldots, n-1 \text{ és } d_n := 1.$$

Ezután ellenőrizni lehet, hogy a Jd vektor koordinátái negatívak, ami megmutatja, hogy J diagonálisan dominált, és ebből következően Hurwitz-típusú mátrix. Ezzel bebizonyítottuk az állítást. □

2. LEMMA. *A (7) rendszernek létezik egyetlen globálisan aszimptotikusan stabilis egyensúlyi helyzete Ω-ban.*

Bizonyítás. Mivel Ω kompakt, konvex, pozitívan invariáns halmaza a (7) rendszernek, tartalmaz legalább egy egyensúlyi helyzetet. Az előző lemma alapján minden egyensúlyi helyzet $\mathrm{int}(\Omega)$-ban van.
A Brouwer-féle fokszám C^1 leképezésekre vonatkozó definíciója :

$$d(F, \mathrm{int}(\Omega), 0) = \sum_i \mathrm{sign} \det J(x_i^*),$$

ahol $J(x_i^*)$ a (7) rendszer Jacobi-mátrixa az egyensúlyi pontban, és a szummázás végigfut az összes egyensúlyi ponton. A (7) rendszerhez tartozó F vektormező Brouwer-féle fokszáma $\mathrm{int}(\Omega)$-ra és 0-ra nézve jól definiált; jelölje $d(F, \mathrm{int}(\Omega), 0)$. Továbbá azt állítjuk, hogy:

$$d(F, \mathrm{int}(\Omega), 0) = (-1)^n.$$

Ahhoz, hogy ezt belássuk, válasszunk tetszőlegesen egy $\tilde{x} \in \mathrm{int}(\Omega)$ pontot, és tekintsük Ω-n a következő vektormezőt:

$$G(x) = \tilde{x} - x.$$

Nyilvánvalóan:

$$d(G, \mathrm{int}(\Omega), 0) = (-1)^n.$$

Megmutatjuk, hogy F és G homotóp, és mivel a Brouwer-féle fokszám topológikusan invariáns, ebből az állításunk következik. Legyen:

$$H(x,t) = tF(x) + (1-t)G(x).$$

Ekkor H folytonos $\Omega \times [0,1]$-en, $H(x,0) = G(x)$ és $H(x,1) = F(x)$. Már csak azt kell bizonyítanunk, hogy $H(x,t) \neq 0$ minden $x \in \partial\Omega$ és minden $t \in (0,1)$ esetén. Indirekt tegyük fel, hogy létezik $\tilde{x} \in \partial\Omega$ és $t \in (0,1)$, amellyel:

$$F(\tilde{x}) = -\frac{1-\tilde{t}}{\tilde{t}} G(\tilde{x}).$$

Ebből következik, hogy F $\tilde{x}$-ben kifelé mutat (míg $G(\tilde{x})$ világosan befelé irányul). De ez ellentmond annak, hogy Ω pozitívan invariáns, és ez így bizonyítja állításunkat. □

Az előző lemma alapján tudjuk, hogy a (7) rendszer Jacobi-mátrixa az egyensúlyi pontokban nem szinguláris, ennélfogva az egyensúlyi pontok száma véges.

Az előző lemmából adódik, hogy minden x_i^* egyensúlyi pont hiperbolikus és lokálisan aszimptotikusan stabilis, azaz:

$$\operatorname{sign} \det J(x_i^*) = (-1)^n,$$

és ebből következik, hogy csak egy egyensúlyi pont lehet.

A globális aszimptotikus stabilitás az 5. lemmából következik, amit a B Függelékben bizonyítunk be. Ahhoz, hogy belássuk, hogy ez az eredmény alkalmazható, vegyük észre először, hogy mivel Ω kompakt és pozitívan invariáns halmaz, a (7) rendszer folytonos félfolyamot generál. A 4. feltétel világos Ω kompaktsága miatt. A 2. feltétel következik abból a tényből, hogy a (7) rendszer kooperatív Ω-n, és ezáltal monoton félfolyamot generál egy olyan rendezéssel, amit a szokásos $\mathbb{R}^n$-beli komponensenkénti rendezést ad.[2] A 3. feltételt most bizonyítottuk, és az 1. feltétel is teljesül. (Bizonyítás: tetszőleges kompakt $K \subset \Omega$ esetén, minden $i = 1, \ldots, n$ mellett legyen $p_i^* \in K$ valamilyen maximális i komponensű pont K-ban. Megjegyezzük, hogy K-ban van ilyen p_i^*, mivel az i -dik komponensre való vetítés folytonos és K kompakt. Ω rács, azaz $\sup(a,b) \in \Omega$, ha $a, b \in \Omega$. Következésképpen $p := \sup_i(p_i^*) \in \Omega$, és könnyű látni, hogy $\sup(K) = p$. Az $\inf(K) \in \Omega$ állítás hasonlóan bizonyítható.)

Megjegyzés: A globális aszimptotikus stabilitást bebizonyíthattuk volna Smillie [26], sőt Mierczynski [22] eredményeinek felhasználásával is. De ezek megkövetelik a folyam egy erősebb monotonitási tulajdonságának ellenőrzését, amit itt most elkerültünk. A Smillie eredményeire támaszkodó bizonyítást arra az esetre, ahol minden komplex csak egy anyagfajtából áll, lásd [5]-ben.

Az *1. tétel bizonyítása:* Következik a (3) rendszer (7) rendszerré való redukciójából és transzformációjából, a 2. lemmával összekapcsolva.

[2]Itt kihasználtuk, hogy Ω konvex, azaz p-konvex. Ez következik a [17] hivatkozás 3.1.1. tételéből és 3.1.1. megjegyzéséből.

3. Diffúzió hozzávétele

A közönséges differenciálegyenletet használó modellek, mint amilyet (3)-ban szemléltünk, implicite felteszik, hogy a reakciók jól kevert környezetben mennek végbe. Bár ez ésszerű feltevés, amikor a diffúzió a reakció időskálájához viszonyítva gyors, nyilvánvalóan érdeke a diffúzió hatását explicite belefoglalni. Ez a (szemilineáris parabolikus néven is ismert) parciális differenciálegyenletekhez; a *reakciódiffúzió*-egyenletekhez vezet.

Ebben a részben megmutatjuk, hogyan kell az eredményünket olyan esetre kiterjeszteni, amikor a diffúzió benne van a modellben. Eredményünk – az $X_1 + X_2 \rightleftharpoons X_3$ reakció speciális példája esetén, a kinetikus tömeghatás törvényét feltételezve – átfedést mutat a [23]-as hivatkozással. Az a cikk a kémiai reakciók Feinberg–Horn–Jackson-féle (FHJ) elméletének közönséges differenciálegyenletről reakciódiffúzió-problémákra való kiterjesztésével foglalkozott (lásd például [11, 19, 29, 7] hivatkozást). (Lásd még a [24] hivatkozást, mely diffúziót is tartalmazó FHJ-rendszerekre vonatkozó konvergenciaeredményeket mutat be, hiányos bizonyítással.) A [24, 23] hivatkozásokban közölt módszerek a Ljapunov-függvényeket veszik alapul, és ebből kifolyólag különböznek a mi megközelítésünktől, amely lehetővé teszi reakciók egy másik osztályának kezelését, és nem kell a kinetikus tömeghatás törvényére szorítkoznunk. Másrészről, számos olyan kémiai reakció van, amelyik FHJ típusú, de nem monoton, és ezért nem lehet a mi módszerünkkel kezelni.

Ebben a szakaszban az a célunk, hogy megmutassuk, hogy a PDE-modellre vonatkozó analóg konvergenciaeredmények hogyan következnek az ODE-kre vonatkozók egyszerű következményeként. (A bizonyítás egy lehetséges alternatívája az lenne, hogy minden eredményt kezdettől fogva a monoton reakció-diffúzió-rendszerek elméletének keretén bizonyítanánk be, de az ODE-kre való redukció jóval egyszerűbb.) Általában a tér és idő $(q,t) \mapsto x(q,t)$ függvényeire vonatkozó PDE-feladatokat tekintünk kezdeti feltételekkel és Neumann-féle (zéró fluxusú) peremfeltétellel, ahol a pont az idő szerinti deriváltat, x_ν a normális irányú deriváltat jelenti, f monoton vektormező, és L a diffúzió parciális differenciáloperátora:

$$\begin{aligned} \dot{x}(q,t) &= (Lx)(q,t) + f(q,x(q,t)) \qquad && t>0,\ q\in Q \\ x_\nu(q,t) &= 0 && t>0,\ q\in \partial Q \\ x(q,0) &= x_0 && q\in \bar{Q}. \end{aligned} \tag{8}$$

Az a kulcsmegfigyelés, amit tenni akarunk (alkalmas technikai feltételek mellett), hogy a (8) rendszer minden megoldása kovergál az egyértelmű homogén egyensúlyi helyzethez: $x(q,t) \to c$, ha $t \to +\infty$, feltéve, hogy a problémához társított $\dot{x} = f \circ x$ ODE minden megoldása c-hez konvergál. Így a korábban bizonyított eredmények kiterjeszthetők a diffúziós esetre (f monotonitása elengedhetetlen – vessük össze a diffúziós instabilitás jelenségével, amely a mintázatképződés aktivátor-inhibitor mechanizmusában merül fel). Először kifejtjük a hátteret a [27] hivatkozás 7. feje-

zetéből a monoton reakció-diffúzió-rendszerekre vonatkozó eredményekre koncentrálva, egyesítve azokat a [1]-ben lévő technikai tényekkel.

A Q halmaz teret reprezentál, korlátos, nyílt, összefüggő részhalmaza az $\mathbb{R}^M$ euklideszi térnek, a (C^4 osztályú) sima ∂Q határral. Az f vektormező kétszer folytonosan differenciálható. Jelölje x_ν a q pontban a $\nu(q)$ külső normális egységvektor irányában a ∂Q halmaz q pontjában vett iránymenti deriváltat. Válasszunk $\mathbb{R}^N$-nek egy X nem üres, zárt részhalmazát, megszorítva a koncentráció megengedett értékeire, amilyen például a nemnegatív ortáns vagy az 1. lemmában használt Ω kompakt és konvex állapottér, és tegyük fel, hogy X pozitívan invariáns az $\dot{x} = f \circ x$ közönséges differenciálegyenletre vonatkozóan (az alábbiakban X-re két kiegészítő feltevést is teszünk). A kezdeti feltétel egy

$$x_0 : \bar{Q} \to X$$

függvény, ami kétszer folytonosan differenciálható és kielégíti az $(x_0)_\nu = 0$ peremfeltételt. A (8) rendszer „megoldásán" értünk egy

$$x = (x_1, \ldots, x_n)^\mathsf{T} : \bar{Q} \times (0,T) \to X$$

függvényt, amire (8) fennáll,

$\frac{\partial x_i}{\partial t}, \frac{\partial x_i}{\partial q_j}, \frac{\partial^2 x_i}{\partial q_j \partial q_k}$ Hölder-folytonosak a $Q \times (0,T)$ halmazon minden i,j,k-ra, és $\frac{\partial x_i}{\partial q_j}, x_i$ folytonos a $\bar{Q} \times (0,T)$ halmazon minden i,j-re.

Ezek a feltevések olyanok, mint az [1] hivatkozásban; [27]-ben viszont csak azt követelik meg, hogy $\frac{\partial x_i}{\partial q_j}$ legyen folytonos a $\bar{Q} \times (0,T)$ tartományon (a Hölder-folytonosságot is az enyhébb folytonossággal helyettesítik), de a kezdeti feltételekre kevesebb regularitási feltételt tesznek.

Az L differenciáloperátor a következő alakú:

$$Lx = (L_1 x_1, \ldots, L_n x_n)^\mathsf{T},$$

ahol minden i-re:

$$L_i = \sum_{j,k=1}^{n} a^i_{j,k}(q) D_j D_k + \sum_{k=1}^{n} a^i_k(q) D_k,$$

$a^i_{kj} = a^i_{jk} \in C^2(\bar{Q})$, valamint L egyenletesen elliptikus, azaz:

$$\exists \mu > 0, \text{ hogy } \xi^\mathsf{T} A_i(q) \xi \geq \mu |\xi^2| \quad \forall \xi \in \mathbb{R}^n, \quad \forall i = 1,2,\ldots,n,$$

ahol $A_i(q) = (a^i_{jk}(q))$. Számunkra az az eset a legfontosabb példa, mikor az anyagfajták diffúziója független egymástól: $a^i_{jj} \equiv d_i > 0$ és $a_{jk} \equiv 0$ minden $j \neq k$-ra, azaz például $L_i x_i = d_i \Delta x_i$, ahol a Δ a Laplace-operátor.

Két kiegészítő feltételt kell tennünk a megengedett állapotvektorok X halmazára. Már megköveteltük, hogy ez legyen invariáns az $\dot{x} = f \circ x$ dinamikára nézve. Egy második feltétel, hogy invariánsnak kellene lennie a diffúzióra nézve is, abban

az értelemben, hogy az $\dot{x} = Lx$, $x_0(q,0) \in X$ kezdeti feltételű lineáris probléma megoldásának minden $q \in Q$ esetén teljesítenie kell, hogy minden $t > 0$, minden $q \in Q$ mellett $x(q,t) \in X$. Tételezzük fel mostantól fogva, hogy *vagy* Q tetszőleges, nyílt, konvex halmaz, azonban minden L_i operátor megegyezik, (például az anyagfajták diffúziója független, és $d_i = d_j$ minden i,j indexre), *vagy* az L_i operátorok tetszőlegesek, azonban a Q halmaz azonos egy (a,b) „téglalappal", ahol $b - a \in \mathbb{R}^n_+$ (lehetséges, hogy a koordinátái között szerepel $-\infty$, és b koordinátái között szerepel $+\infty$ is).

Mostantól feltételezzük, hogy adott egy rendezés $\mathbb{R}^n$-ben. Az utolsó feltétel hálófeltétel az X halmazon (lásd még a B Függelékben): bármely $S \subseteq K$ kompakt részhalmazra, mind $\inf(S)$, mind $\sup(S)$ definiálva van, és X-hez tartozik. Azt mondjuk, hogy a vektormező *kvázi-monoton* (az $X \subseteq \mathbb{R}^n$-n adott rendezésre nézve), ha $\dot{x} = f \circ x$ folyama monoton. Ha adott az x és y X-beli értékeiket felvevő függvény, akkor azt írjuk, hogy $x \preceq y$, ha $x(q,t) \preceq y(q,t)$ minden olyan (q,t) esetén, amely az értelmezési tartományuk közös eleme. A következő egy változata a [27] hivatkozás 3.4. tételének. Ezt specializáltuk a PDE esetre (a kézikönyvben általánosabban parciális differenciálegyenlőtlenségekre van megadva), és tetszőleges rendezésekre mondtuk ki. (A könyvben az állítás csak kooperatív rendszerekre vonatkozik, de hasonló bizonyítás érvényes tetszőleges rendezésre is, lásd a 142. oldalon.)

2. TÉTEL. *Ha f kvázi-monoton, és az y, z megoldások a $[0,T)$ intervallumon vannak értelmezve úgy, hogy $y(\cdot,0) \preceq x_0 \preceq z(\cdot,0)$ $\bar{Q}$-n, akkor a (8) rendszernek létezik egyetlen x megoldása, és az értelmezve van legalább a $[0,T)$ intervallumon, és $y \preceq x \preceq z$ a $\bar{Q} \times [0,T)$-n.*

Megtettük az előkészületeket arra, hogy kimondjuk következtetéseinket. Az első megállapításunk a következő:

3. TÉTEL. *Tegyük fel, hogy f kvázi-monoton, és létezik $\xi \in X$, hogy $\dot{x} = f \circ x$, $x_0 \in X$ minden megoldása $t \to +\infty$ esetén ξ-hez konvergál. Ekkor a (8) rendszernek létetik egyetlen $x(q,t)$ megoldása, amely értelmezve van minden $t > 0$-ra, és minden x_0 kezdeti feltételre létezik $x(q,t) \to \xi$ egyenletesen, ha $t \to +\infty$, $q \in Q$ mellett.*

Bizonyítás. Az állítás bizonyításához először legyen $y : \bar{Q} \to X$ olyan függvény, amely állandó: megegyezik x_0 minimumával, és $z : \bar{Q} \to X$ pedig olyan függvény, mely állandó: megegyezik x_0 maximumával. Továbbá, megjegyezzük, hogy az $\dot{x} = f \circ x$, $x(0) = y$ kezdetiérték-probléma $y(t)$ megoldása (mely minden t-re értelmezve van, és $t \to +\infty$ esetén ξ-hez konvergál) a (8) rendszernek is megoldása, egyszerűen az $y(q,t) := y(t)$ definícióval. Hasonlóan z-re, és a 2. tétel esetéhez jutottunk. Alkalmazva ezt a 2. tételt a $[0,T)$ növekvő, véges intervallumaira, $x(q,t)$ egzisztenciáját és unicitását kapjuk a $[0,+\infty)$ intervallumon. Továbbá, $y(q,t) \preceq x(q,t) \preceq z(q,t)$, valamint $y(q,t) \to \xi$ és $z(q,t) \to \xi$ (q-ban egyenletesen), ami a következtetést adja. □

Sajnálatosan – bármilyen elegáns is a 3. tétel – ez nem elégséges önmagában, amikor az eredeti (3) rendszerrel foglalkozunk, mert ennek a rendszernek sok egyensúlya van. Kiegészítő feltételt kell tennünk, mégpedig azt, hogy minden diffúziós együttható megegyezik.

4. TÉTEL. *Tegyük fel, hogy f olyan, mint az 1. tételben, és valamilyen $d > 0$-ra $L_i x_i = d\Delta x_i$ az állapottér minden koordinátájára. Ekkor a (8) rendszer minden megoldása konvergál a (homogén) stacionárius állapothoz.*

Bizonyítás. A bizonyításhoz használjuk fel ugyanazt a koordináta-transzformációt, mint korábban. A PDE-re alkalmazva, ez a következő formáját eredményezi az egyenletnek:

$$\begin{aligned}
\dot{z_1} &= -r_1(a_1^1 z_1) + r_{-1}(a_2^1(z_2 - z_1)) + d\Delta z_1 \\
&\vdots \\
\dot{z_k} &= -r_k(a_k^1(z_k - z_{k-1})) + r_{-k}(a_{k+1}^1(z_{k+1} - z_k)) + d\Delta z_k \quad k = 2, \ldots, n-1 \\
&\vdots \\
\dot{z_n} &= -r_n(a_n^1(z_n - z_{n-1})) + r_{-n}(a_{n+1}^1(C - z_n)) + d\Delta z_n.
\end{aligned}$$

A 2. lemma és a 3. tétel összevetéséből ismert, hogy ennek a rendszernek minden megoldása konvergál egy (egyértelmű) homogén stacionárius állapothoz. Így az $y_i = x_i^1$ változók szintén konvergálnak egy bizonyos stacionárius állapothoz. Most bebizonyítjuk, hogy a további változók szintén konvergálnak.

Emlékezzünk vissza, hogy a közönséges differenciálegyenletnek (amikor nincs diffúzió) létezik $n_i - 1$ független lineáris első integrálja, amint (5)-ben megmutattuk:

$$\dot{Z}_{ik} = 0, \quad \forall i, \forall k = 2, \ldots, n_i,$$

ahol $Z_{ik} = x_i^k/a_i^k - x_i^1/a_i^1$. Ebből ugyanazt a kifejezést kapjuk, mint (6)-ban:

$$x_i^k(t) = \beta_i^k x_i^1(t) + \alpha_i^k, \qquad \forall i, \forall k = 2, \ldots, n_i$$

valamilyen $\alpha_i^k \in \mathbb{R}$ (ami a kezdeti feltételektől függ) és $\beta_i^k > 0$ számokkal. Így, amikor x_i^1 konvergens, ugyanezt a következtetést vonhatjuk le a többi x_i^k változóra is. Amikor hozzávesszük a diffúziót, akkor ez az érvelés nem érvényes. Az (5) egyenlet ehelyett:

$$\dot{Z}_{ik} = LZ_{ik}, \quad \forall i \forall k = 2, \ldots, n_i,$$

alakúvá válik, ahol $LZ = d\Delta Z$ azzal a Neumann-feltétellel, hogy a határpontokban $(Z_{ik})_\nu = 0$. Ennek a PDE-nek minden megoldása konstanshoz konvergál, a kezdeti értékek az $\frac{1}{|Q|}\int_Q Z_{ik}(q,0)dq$ átlagához, ahol $|Q|$ Q-nak a mértéke. (Vázlatos bizonyítás: Létezik a Neumann–Laplace-féle önadjungált operátornak λ_i sajátértékeiből és a hozzájuk tartozó Φ_i, $i = 1, 2, \ldots$ sajátvektoraiból álló olyan sorozat,

amelynek tagjai megoldásai az $L\Phi + \lambda\Phi = 0$, $\Phi_\nu = 0$ problémának. Ezekre fennáll, hogy $\lambda_1 = 0$, $\Phi_1 = 1$, és $\lambda_i > 0$ minden $i >$ 1-re, valamint Φ_i ortogonális bázisa az L^2 térnek. Most vegyünk egy tetszőleges folytonos és korlátos x_0 kezdeti feltételt, és tekintsük ezt az L^2 tér elemének, és fejtsük ki a bázis szerint: $x(q,0) = \sum_{i-1}^{\infty} b_i\Phi_i(q)$, ekkor a $\dot{Z} = LZ$ rendszernek $x(q,t) = \sum_{i-1}^{\infty} b_i e^{-\lambda_i t}\Phi_i(q)$ megoldása az $x(q,0) = \sum_{i-1}^{\infty} b_i\Phi_i(q)$ kezdeti feltétellel, és ez L^2-ben a b_1 első Fourier-együtthatóhoz konvergál, ami a megkövetelt átlag.) Összefoglalva, mind $x_i^k/a_i^k - x_i^1/a_i^1$, mind x_i^1 egy számhoz konvergál, és ugyanígy minden x_i^k is. □

A Függelék: A diagonálisan dominált mátrixok Hurwitz-típusú mátrixok

Ez egy jól ismert eredmény (lásd például [25]). Egy rövid, Gersgorin-tételére alapozott bizonyítását adjuk. Először egy speciális esetet vizsgálunk, és utána megmutatjuk, hogy ebből mindig levezethető az általános eset. Ha az A mátrix diagonálisan dominált a $d_i = 1$ ($\forall i = 1, \ldots, n$) számokra nézve, azaz:

$$a_{ii} + \sum_{j \neq i} |a_{ij}| < 0, \quad \forall i = 1, \ldots, n,$$

akkor a Gersgorin-tételből következik, hogy A Hurwitz-típusú mátrix.

Ha az A mátrix diagonálisan dominált olyan pozitív d_i számok halmazára nézve, amelynek nem mindegyike 1, akkor megmutatjuk, hogy az A mátrix hasonló az A^* mátrixszal, ami diagonálisan dominált a $d_i = 1(\forall i = 1, \ldots, n)$ esetén. Az eredmény ebből egyenesen következik.

Hogy bebizonyítsuk ezt az állítást, definiáljuk a T diagonális mátrixot a következő koordinátákkal:

$$t_{ii} = 1/d_i, \quad \forall i = 1, \ldots, n.$$

Ezután egy egyszerű számolással megmutatható, hogy $A^* = TAT^{-1}$ úgy, hogy $a_{ij}^* = a_{ij}d_j/d_i \quad \forall i, j = 1, \ldots, n$ esetén. De másfelől ebből következik, hogy:

$$a_{ii} + \sum_{j \neq i} |a_{ij}^*| = (d_i a_{ii} + \sum_{j \neq i} |a_{ij}|\, d_j)/d_i.$$

Az n számú mennyiség mindegyike negatív, az állításunk bizonyítása befejeződött.

B függelék: Egy egyértelmű egyensúllyal bíró monoton folyamokra vonatkozó globális attraktivitási eredmény

Tekintsük az X metrikus teret a d metrikával, és tegyük fel, hogy adott az X téren egy $\preceq$ parciális rendezés. Feltételezzük, hogy a parciális rendezés és az X tér

metrikus topológiája kompatibilis a következő értelemben: ha $x_n \to x$ és $y_n \to y$ konvergens sorozat X-ben, és ha $x_n \preceq y_n$, akkor $x \preceq y$. Időnként a szabálytalan $x \preceq A$ (valamilyen $x \in X$ és $A \subset X$ mellett) írásmódot alkalmazzuk, ami a $\forall y \in A$-ra $x \preceq y$ állítást jelöli. Alkalmazni fogjuk az ismert rendezéselméleti fogalmakat: $\sup(A)$ jelöli a legkisebb felső, illetve $\inf(A)$ a legnagyobb alsó határát az $A \subset X$ halmaznak, feltéve, hogy léteznek ilyenek X-ben. Ha $p, q \in X$ és $p \preceq q$, definiáljuk a rendezés $[p, q] := \{x \in X | p \preceq x \preceq q\}$ intervallumát. Egy $A \subset X$ halmazt a rendezésre nézve *konvexnek* nevezünk, ha $[p, q] \subset A$ minden olyan $p, q \in A$ pontpárra, amelyre $p \preceq q$.

Az X halmazon vett Φ folytonos félfolyam által generált dinamikával fogunk foglalkozni. Emlékezzünk rá, hogy ez folytonos $\Phi : \mathbb{R}_+ \times X \to X$ leképezés, $(\Phi_t(x) := \Phi(t, x))$, amelyre $\Phi_0 = \mathrm{Id}$, és $\Phi_t \circ \Phi_s = \Phi_{t+s}$ minden $t, s \in \mathbb{R}_+$ mellett.

Az X térre és a Φ leképezésre a következő feltételeket adjuk:

1. X minden C kompakt részhalmazára igaz, hogy $\inf(C)$, $\sup(C) \in X$.
2. Φ monoton a $\preceq$ rendezésre nézve, azaz (1) fennáll.
3. Φ-nek X-ben létezik az a egyértelmű egyensúlyi pontja.
4. Minden $x \in X$-re az $O(x) := \{\Phi_t(x) | t \in \mathbb{R}_+\}$ pálya lezártja kompakt X-ben.

Az utolsó, 4. feltétel magában foglalja nevezetesen azt, hogy x ω-határhalmaza – jelölje ezt $\omega(x)$ – nem üres, konvex, invariáns halmaz (ez azt jelenti, hogy $\Phi_t(\omega(x)) = \omega(x)$ minden $t \in \mathbb{R}_+$ esetén), és $\lim_{t \to +\infty} d(\Phi_t(x), \omega(x)) = 0$ (ahol az $x \in X$ pont és az $A \subset X$ halmaz távolsága a szokásos $d(x, A) = \inf_{y \in A} d(x, y)$). Az 1–4. feltételekből kapjuk a következő állítást:

5. TÉTEL. *Az a egyensúlyi pont globálisan attraktív a Φ leképezésre nézve.*

Bizonyítás. Vegyünk egy $x \in X$ pontot, és tekintsük $\omega(x)$-et. Ekkor definiálni tudjuk az:

$$m := \inf(\omega(x)) \text{ és az } M := \sup(\omega(x))$$

értékeket. Azt állítjuk, hogy:

$$\Phi_t(x) \preceq m, \quad \forall t \in \mathbb{R}_+. \tag{9}$$

Ahhoz, hogy ezt belássuk, be fogjuk bizonyítani, hogy minden $t \geq 0$-ra $\Phi_t(m) \preceq \omega(x)$, amiből (9) következni fog, ha $\omega(x)$-nek m a legnagyobb alsó határa.

Válasszunk egy $t \geq 0$ időpontot, és egy tetszőleges $p \in \omega(x)$ pontot. Meg kell mutatnunk, hogy $\Phi_t(m) \preceq p$. $\omega(x)$ invarianciájából következik, hogy valamilyen $q \in \omega(x)$ esetén $\Phi_t(q) = p$. Mivel $q \in \omega(x)$ ezért $m \preceq q$, és ennélfogva a monotonitásból következik, hogy $\Phi_t(m) \preceq \Phi_t(q) = p$, tehát ezzel bebizonyítottuk (9)-et.

A monotonitásból következik, hogy $\Phi_t(m)$ csökkenő, azaz $0 \leq t_1 \leq t_2$ esetén $\Phi_{t_2}(m) \preceq \Phi_{t_1}(m)$. (Egyszerűen (9)-re alkalmazzuk $\Phi_{t_1}(m)$-t, ahol $t = t_1 - t_2$.)

Most azt állítjuk, hogy $\omega(x) = \{a\}$.[3] Először megmutatjuk, hogy ha $p, q \in \omega(m)$, ebből következik, hogy $p = q$. Válasszunk $\Phi_{t_k}(m) \to p$ és $\Phi_{t_l}(m) \to q$ sorozatokat $(t_k, t_l \to +\infty)$. Mivel $\Phi_t(m)$ nemnövekvő, ezért lehetséges, hogy találunk minden t_k-hoz valamilyen $t_{l(k)} \geq t_k$ sorozatot, hogy $\{t_{l(k)}\}$ egy részsorozatát alkotja $\{t_l\}$-nek és $\Phi_{t_{l(k)}}(m) \preceq \Phi_{t_k}(m)$. A határértékek meghatározása után azt kapjuk, hogy $q \preceq p$. Hasonló érveléssel megmutathatjuk, hogy $p \preceq q$, következésképpen $p = q$. Ez azt mutatja, hogy $\omega(m)$ egyelemű. Az ω-határhalmazok invariánsak, és ebből következik, hogy $\omega(m)$-nek tartalmaznia kell egy egyensúlyt. Egyetlen a egyensúlyi pont létezik, és ebből következik, hogy $\omega(m) = a$, ezzel bebizonyítottuk az állítást.

Hasonló érveléssel belátható, hogy $\Phi_t(M)$ monoton növekedő, és így

$$\omega(M) = \{a\}.$$

Végül, minden $t \geq 0$-ra kapjuk, hogy:

$$\Phi_t(m) \preceq m \preceq \omega(x) \preceq M \preceq \Phi_t(M),$$

és $t \to +\infty$ határértéket véve $\omega(x) = a$, és ezzel bebizonyítottuk a tételt. □

Megjegyzés: Ebben a megjegyzésben az első feltétel egy ellenőrzését adjuk abban az esetben, amikor az X állapottér egy részhalmaza a véges dimenziós térnek.

Tegyük fel, hogy az Y egy véges dimenziós normált vektortér, és az X állapottér részhalmaza az Y-nak. Feltételezzük, hogy a $\preceq$ parciális rendezés Y-ban – és egyben X-ben – a $K \subset Y$ kúp által generált. Megjegyezzük, hogy a K kúpot *normális*nak nevezzük, ha $k > 0$-ra $x, y \in Y$, és $0 \preceq x \preceq y$ esetén $|x| \leq k\,|y|$. Könnyű belátni, hogy ha K normális, akkor minden rendezési intervallum egy zárt halmaz Y-ban.

A következő lemma megmutatja, hogy a K kúp az Y *véges-dimenziójú* térben mindig normál.

3. LEMMA. *Legyen Y véges-dimenziójú vektortér a K kúppal, és Y-ban adva egy $\preceq$ parciális rendezés. Ekkor K normális.*

Bizonyítás. Megmutatjuk, hogy:

$$M := \sup\{|z| \quad | \quad 0 \preceq z \preceq x, |x| = 1\}$$

véges, valós szám. Az állításunk következik abból, hogy a normalitás definíciójában szereplő k-t M-nek választjuk. Valóban, $0 \preceq x \preceq y, y \neq 0$ esetén (ezt feltételezhetjük az általánosság megszorítása nélkül) következik, hogy $0 \preceq x/|y| \preceq y/|y|$.

De abból, hogy $|x/|y|\,| \leq M$ és $k = M$, az állításunk következik.

Most bizonyítsuk, hogy M véges. Tegyük fel, hogy nem az, akkor $\{x_n\}$ és $\{z_n\}$ sorozatok kielégítik a $0 \preceq z_n \preceq x_n$ feltételt, hogy $|x_n| = 1$ minden n-re, és

[3] A monoton rendszerek konvergencia kritériumából közvetlenül következik ez az állítás ([27] hivatkozás 1.2.1. tétele), felhasználva, hogy az a egyensúly egyértelmű. Habár itt inkább önálló rövid bizonyítást adunk anélkül, hogy felhasználnánk a monoton rendszerek elméletének eredményeit.

$|z_n| \to \infty$. Tekintsük az $\{y_n\}$ sorozatot, ahol $y_n = z_n/|z_n|$. Nyilvánvalóan, Y egységgömbjének kompaktsága miatt $|y_n| = 1$ minden n-re, és (mivel Y végesdimenziós) tekinthetjük az y_{n_k} részsorozatot, ami y^*-hoz konvergál. Világos, hogy $|y^*| = 1$ és $x_n/|z_n| \to 0$. Így a parciális rendezés kompaktsága és a metrikus topológia által kapjuk, hogy:

$$0 \preceq y^* \preceq 0.$$

De ez ekvivalens azzal az állítással, hogy $y^* \in K \cap (-K)$, és így $y^* = 0$ (mivel K egy kúp). Ez ellentmond azzal, hogy $|y^*| = 1$, és ez bizonyítja állításunkat. □

Megjegyezzük, hogy az X parciális rendezés halmaza egy *háló*, ha $\sup(p,q)$, $\inf(p,q) \in X$ minden $p,q \in X$ esetén. Azt mondjuk, hogy az $S \subset X$ *korlátos rendezés* az X halmazon, ha $a,b \in X$ úgy, hogy $S \subset [a,b]$.

4. LEMMA. *Legyen Y véges dimenziós normált vektortér egy K kúppal, és legyen $X \subset Y$ egy háló. Feltételezzük, hogy X-ben minden korlátos halmaz X-ben korlátos rendezés. Ha C kompakt részhalmaza X-nek, akkor* $\inf(C), \sup(C) \in X$.

Bizonyítás. Csak azt az állítást bizonyítjuk, hogy $\sup(C) \in X$. A bizonyítás hasonló az $\inf(C) \in X$ állításra is.

Mivel C korlátos, így rendezésre nézve is korlátos, és $a,b \in X$, hogy $C \subset [a,b]$. Metrikus térben kompakt halmazok szeparábilisak, így kiválaszthatjuk C-nek egy megszámlálható és sűrű $\{c_k\}$ részhalmazát. Mivel X háló, X-nek képezhetjük egy $\{x_k\}$ sorozatát a következőképpen:

$$x_1 = c_1,$$
$$x_k = \sup(c_k, x_{k-1}), \quad k > 1.$$

Ez a sorozat a következő tulajdonságokkal rendelkezik:

1. $\{x_k\}$ növekvő, azaz $x_k \preceq x_{k+1}$.
2. $\{x_k\} \subset [a,b]$.

Az $[a,b]$ rendezési intervallum zárt (a metrikus topológia és parciális rendezés kompatibilitása által) és zárt (mivel K normális), ezáltal kompakt. Így $\{x_k\}$ növekvő sorozat, amelyik az $[a,b]$ kompakt halmazban marad, és így valamely $x \in [a,b] \subset X$, hogy $x_k \to x$ (ezt az állítást bebizonyítottuk az 5. tétel bizonyításában). Most azt állítjuk, hogy:

$$\sup(C) = x.$$

Két lépésben bizonyítjuk ezt az állítást. Először megmutatjuk, hogy x felső határa C-nek. Másodszor megmutatjuk, hogy ez a legkisebb felső határa.

Választunk egy $c \in C$ elemet. Mivel $\{c_k\}$ sűrű C-ben, a $\{c_k\}$ sorozatból kiválaszthatunk egy $\{c_{n_k}\}$ konvergens részsorozatot c határértékkel. Már ismert, hogy $c_{n_k} \preceq x$ minden n_k-ra, így a kompatibilitásból következik, hogy $c \preceq x$. Végül, legyen y egy tetszőleges felső határa C-nek. Ekkor $c_k \preceq y$ minden k-ra, mivel

$x_k \preceq y$ minden k-ra. Vegyük a határértéket, és alkalmazzuk a kompatibilitást még egyszer, kapjuk, hogy $x \preceq y$, és így x a legkisebb felső határa C-nek. □

Például tegyük fel, hogy $Y = \mathbb{R}^n$ és $K = \mathbb{R}^n_+$, és X vagy $\mathbb{R}^n_+$, vagy $\mathbb{R}^n$. Világos, hogy X egy rács, és az X-ben minden korlátos halmaz rendezés korlátos. Ezentúl, a 4. lemmából következik, hogy az X kompakt részhalmazainak a suprémuma és infinuma X-ben van.

Megjegyzés: Az 5. tétel alkalmazható végtelen dimenziójú téren értelmezett folyamokra is. Például a késleltetett egyenletekben gyakran tekintjük a kompakt intervallumon értelmezett folytonos függvények terét, amilyenek $X = C([-r, 0], \mathbb{R}^n)$ vagy $X = C([-r, 0], \mathbb{R}^n_+)$, a szuprémum norma által indukált, szokásos metrikával, és az $f_1 \preceq f_2$ parciális rendezéssel, amit az $f_2(t) - f_1(t) \in \mathbb{R}^n_+$ minden $t \in [-r, 0]$-ra feltétel definiál. Mindkét esetben a kompakt halmazoknak létezik X-ben infinuma és szuprémuma, lásd a [18] hivatkozásban.

Megjegyzés: A [20] hivatkozásban – aminek a gondolatmenetét itt követjük – jelent meg az 1. feltétel. Sőt ez a feltétel előkerül a [18]-as munkában is. Bár ott a félfolyamokra egy erősebb monotonitási tulajdonságot írnak elő, mégis az egyensúly nem egyértelmű. Az eredmény az, hogy a kvázikonvergens pontok halmaza (egy pont kvázikonvergens, ha az egyensúlyi halmaz tartalmazza a pont határhalmazát) tartalmaz egy nyílt és sűrű halmazt. A bizonyítás a monoton dinamikai rendszerek elméletének számos alapvető eredményére épül.

Habár a kémiai reakcióhálózatokban elért legfontosabb eredményeink bizonyításához az 5. tétel szükséges (1. tétel), az a egyensúlyi pont stabilitásáról általánosabb következtetéseket tudunk levonni, feltéve, hogy az X tér és a Φ folyam további feltételeket is kielégít.

Minden $x \in X$ pont minden környezete tartalmazza x-nek egy kompakt, rendezéskonvex C környezetét.

A következő eredményre jutottunk:

5. LEMMA. *Tegyük fel, hogy minden $t \in \mathbb{R}_+$ értékre Φ_t nyílt leképezés. Az 1-4. feltételek és* C *miatt Φ-nek az a egyensúlyi pontja globálisan aszimptotikusan stabilis.*

Bizonyítás. Az 5. tétel miatt elegendő azt bebizonyítani, hogy a stabilis egyensúly. Ismételten használni fogjuk azt a tényt, hogy minden olyan $p, q \in X$ esetén, amelyre $p \preceq q$ fennáll, arra:

$$\Phi_t([p, q]) \subset [\Phi_t(p), \Phi_t(q)], \quad \forall t \in \mathbb{R}_+$$

is teljesül Φ monotonitása miatt.

Válasszuk a-nak egy tetszőleges U környezetét. A C feltétel miatt a-nak létezik egy kompakt, rendezéskonvex C környezete, hogy $C \subset U$. Az 1. feltétel miatt definiálhatjuk az

$$i := \inf(C) \text{ és } s := \sup(C)$$

értékeket, és tekinthetjük a rendezés $[i, s]$ intervallumát. Ekkor nyilvánvalóan $C \subset [i, s]$, így a-nak $[i, s]$ is egy környezete. Következésképpen Φ_t nyílt hozzárendelés minden $t \in \mathbb{R}_+$ esetén, és $\Phi_t([i, s])$ is környezete a-nak.

Most válasszunk egy $T > 0$ számot úgy, hogy:

$$\Phi_t(i), \Phi_t(s) \in C, \quad \forall t \geq T. \tag{10}$$

Az 5. tétel alapján ilyen T létezik.

Most tekintsük a-nak a $V := \Phi_T([i, s])$ környezetét. Ekkor minden $t \geq 0$-ra kapjuk, hogy:

$$\Phi_t(V) = \Phi_t(\Phi_T([i, s])) \subset \Phi_t([\Phi_T(i), \Phi_T(s)]) \subset [\Phi_{t+T}(i), \Phi_{t+T}(s)] \subset C \subset U,$$

ahol alkalmaztuk a fenti tényt az első két tartalmazásnál; (11)-et és azt, hogy C konvex a rendezésre nézve, a harmadik tartalmazásnál. Ezzel befejeztük a bizonyítást. □

Hivatkozások

[1] H. Amann: *"Invariant sets and existence theorems for semilinear parabolic and elliptic systems,"* J. Math. Anal. Appl. **65**(1978): 432–467.

[2] D. Angeli and E. D. Sontag: *Monotone control systems*, Trans. Autom. Contr. **48**, 1684–1698 (2003).

[3] D. Angeli and E. D. Sontag: *Multistability in monotone I/O systems*, Systems and Control Lett. **51**, 185–202 (2004).

[4] D. Angeli, P. De Leenheer and E. D. Sontag: *A small-gain theorem for almost global convergence of monotone systems*, to appear in Systems Contol Lett.

[5] D. Angeli and E. D. Sontag: *Interconnections of monotone systems with steady-state characteristics*, in Optimal Control, Stabilization, and Nonsmooth Analysis, Eds: de Queiroz, M., M. Maliso-, and P. Wolenski, Springer-Verlag, Heidelberg, 2004, pp. 135–154.

[6] D. Angeli, J. E. Ferrell, Jr., and E. D. Sontag: *Detection of multi-stability, bifurcations, and hysteresis in a large class of biological positive-feedback systems*, Proceedings of the National Academy of Sciences USA **101**, 1822–1827 (2004).

[7] M. Chaves and E. D. Sontag: *"State-estimators for chemical reaction networks of Feinberg- Horn-Jackson zero-deficiency type"*, European J. Control (2002)**8**:343–359.

[8] P. De Leenheer, D. Angeli and E. D. Sontag: *On predator-prey systems and small-gain theorems*, submitted (Preliminary version entitled 'Small-gain theorems for predator-prey systems' has appeared in the Lecture Notes in Control and Information Sciences, **294**, 191–198 (2003)).

[9] P. De Leenheer, D. Angeli and E. D. Sontag: *Crowding effects promote coexistence in the chemostat*, submitted (also DIMACS Tech report 2003-44; preliminary version entitled 'A feedback perspective for chemostat models with crowding effects' has appeared in the Lecture Notes in Control and Information Sciences, **294**, 167–174 (2003).

[10] P. De Leenheer, S. A. Levin, E. D. Sontag and C. A. Klausmeier: *Global stability in a chemostat with multiple nutrients*, submitted (also DIMACS Tech report 2003-40).

[11] M. Feinberg: *"Chemical reaction network structure and the stabiliy of complex isothermal reactors - I. The deficiency zero and deficiency one theorems"*, Review Article **25**, Chemical Eng. Science (1987)**42**:2229–2268.

[12] M. W. Hirsch: *Systems of differential equations which are competitive or cooperative I: limit sets*, SIAM J. Appl. Math. **13**, 167–179 (1982).

[13] M. W. Hirsch: *Systems of differential equations which are competitive or cooperative II: convergence almost everywhere*. SIAM J. Math. Anal. **16**, 423–439 (1985).

[14] M. W. Hirsch: *Systems of differential equations which are competitive or cooperative III: competing species*, Nonlinearity **1**, 51–71 (1988).

[15] M. W. Hirsch: *Systems of differential equations which are competitive or cooperative IV: Structural stability in three dimensional systems*, SIAM J. Math. Anal. **21**, 1225–1234 (1990).

[16] M. W. Hirsch: *Systems of differential equations which are competitive or cooperative V: Convergence in 3-dimensional systems*, J. Diff. Eqns. **80**, 94–106 (1989).

[17] M. W. Hirsch and H. L. Smith: *Competitive and cooperative systems: a mini-review*, Lecture Notes in Control and Information Sciences, **294**, 183–190 (2003).

[18] M. W. Hirsch and H. L. Smith: *Generic quasi-convergence for strongly order preserving semi- ows: a new approach*, preprint.

[19] F. J. M. Horn, and R. Jackson: *"General mass action kinetics"*, Archive for Rational Mechanics and Analysis (1972)**49**:81–116.

[20] J. F. Jiang: *On the global stability of cooperative systems*, Bull. London Math. Soc. **26**, 455–458 (1994).

[21] H. Kunze and D. Siegel: *Monotonicity properties of chemical reactions with a single initial bimolecular step*, J. Math. Chem. **31**, 339–344 (2002).

[22] J. Mierczynski: *Strictly cooperative systems with a first integral*, SIAM J. Math. Anal. **18**, 642–646 (1987).

[23] M. Mincheva, D. Siegel: *"Stability of mass action reaction diffusion systems"*, Nonlinear Analysis **56**(2004): 1105–1131. 13

[24] A. J. Shapiro: *"The statics and dynamics of multicell reaction systems"*, Ph.D. Thesis, The University of Rochester, 1975, 176pp. (http://wwwlib.umi.com/dissertations/fullcit/7614785)

[25] D. D. Siljak: *Large-scale dynamic systems*, Elsevier North-Holland, 1978.

[26] J. Smillie: *Competitive and cooperative tridiagonal systems of differential equations*, SIAM J. Math. Anal. **15**, 530–534 (1984).

[27] H. L. SMITH: *Monotone Dynamical Systems*, AMS, Providence, 1995.

[28] H. L. SMITH AND P. WALTMAN: *The Theory of the Chemostat*, Cambridge University Press, Cambridge, 1995.

[29] E. D. SONTAG: *"Structure and stability of certain chemical networks and applications to the kinetic proofreading model of T-cell receptor signal transduction"*, IEEE Trans. Automatic Control (2001)**46**:1028–1047. Errata in IEEE Trans. Automatic Control (2002)**47**:705.

[30] A. I. VOLPERT, V. A. VOLPERT AND V. A. VOLPERT: *Traveling wave solutions of parabolic systems* (AMS, Providence, 1994)

[31] S. WALCHER: *On cooperative systems with respect to arbitrary orderings*, J. Math. Anal. Appl. **263**, 543–554 (2001).

PATRICK DE LEENHEER
Department of Mathematics, University of Florida
411 Little Hall, PO Box 118105
Gainesville, FL 32611–8105
USA
deleenhe@math.ufl.edu

DAVID ANGELI
Dip. di Sistemi e Informatica Università di Firenze
Via di S. Marta 3, 50139 Firenze
Italy

EDUARDO D. SONTAG
Department of Mathematics, Rutgers University
New Brunswick, NJ 08903
USA

VÁRDAI JUDIT
SZIE, Ybl Miklós Építéstudományi Kar
Budapest, 1146 Thököly út 74.
Vardai.Judit@ybl.szie.hu

MONOTONE CHEMICAL REACTION NETWORKS

PATRICK DE LEENHEER, DAVID ANGELI, EDUARDO D. SONTAG, JUDIT VÁRDAI

We analyze certain chemical reaction networks and show that every solution converges to some steady state. The reaction kinetics are assumed to be monotone but otherwise arbitrary. When diffusion effects are taken into account, the conclusions remain unchanged. The main tools used in our analysis come from the theory of monotone dynamical systems. We review some of the features of this theory and provide a selfcontained proof of a particular attractivity result which is used in proving our main result.

Alkalmazott Matematikai Lapok **26** (2009), 403-404.

SZAKMÁNK TÖRTÉNETE

(A felelős szerkesztő megjegyzése)

A Kedves Olvasó talán észrevette, hogy lapunk igyekszik megemlékezni minden elhunyt kollégánkról. Ezzel nem csupán az eltávozottak előtti tisztelgés a célunk, hanem az is, hogy szakmánk történetét, pontosabban annak egy részét megőrizzük. Egy szakma annyira becsüli meg saját magát, amennyire nem hagyja múltját elfelejteni.

Mostanában két munka is a kezembe akadt, amelyik az operációkutatás történetével foglalkozik különböző szempontból. [3] egyszerűn durva történelemhamisítás. Magyar szempontból itt csak annyit érdemes megjegyezni, hogy szándékosan nem ír az operációkutatáson belüli német-német kapcsolatokról. Így természetesen szóba se jön, hogy megemlítse ezen kapcsolatok kialakításában és fenntartásában oly fontos szerepet játszó mátrafüredi konferenciák sorozatát. [2] értékes mű, nagyon sok adatot tartalmaz, de a válogatás szubjektív. Egy olyan nagy területet, mint a sztochasztikus programozás, lényegében elintéz egyetlen cikkel, nevezetesen Dantzig és Beale eredeti dolgozatával [1]. Ezért, ha valaki meg akarná írni az operációkutatás rendszeresen kifejtett történetét, [2]-ből sok adatot meríthetne, de minden területnek még alaposan utána kellene néznie.

Azt mondhatjuk tehát, hogy még nemzetközi szinten is gond van a szakma történetének megírásával. Sajnos a hazai alkalmazott matematikát illetően még rosszabbul állunk. A teljesség igénye nélkül néhány megírandó téma: a biztosítási matematika kezdetei a II. Világháború előtt, benne Vincze István munkássága; a már említett mátrafüredi konferenciák története; az első számítógépek beszerzése és használata; az INFELOR, a NIMIGÜSZI, az SZKI alapítása és működése különkülön; a Rényi Intézet elődjének hozzájárulása az alkalmazott matematikához; az operációkutatás kezdetei hazánkban. A sort hosszan lehetne még folytatni. Elnézést kérek azoktól, akiknek kedves témája kimaradt. De minden terület egyaránt fontos.

Az Alkalmazott Matematikai Lapok szívesen közölne a jövőben egy-egy terület történetével foglalkozó dolgozatokat. Ezek írására bíztatom a Kedves Olvasót.

VIZVÁRI BÉLA

Hivatkozások

[1] G.B. DANTZIG, E.M.L. BEALE: *Linear programming under uncertainty* Management Science, 1(1955), 197–206.

[2] SAUL I. GASS, ARJANG A. ASSAD: *An Annotated Timeline of Operations Research (An Informal History)*, Kluwer Academic Publisher, 2005, ISBN 1-4020-8116-2.

[3] WOLFGANG LASSMANN, DIETER EHRENBERG, ROLF ROGGE, WALTER RUNGE, PETER STAHLKNECHT: *40 Years of oOperations Research (OR) in the GDR (1949–1989)*, OR News (The Magazine of GOR), **No. 36**, June 2009, 9–12.

HELYESBÍTÉS

Sajnálatos módon az előző szám tartalomjegyzékében az egyik Szerző, Baranyi Lászó neve kétszer is hibásan szerepelt.

Helyesen:

illetve

Az Alkalmazott Matematikai Lapok szerkesztőbizottsága nevében mind a tisztelt Szerzőtől, mind az Olvasóktól elnézést kérünk.

Az Alkalmazott Matematikai Lapok megjelenését támogatja
a Magyar Tudományos Akadémia Könyv- és Folyóiratkiadó Bizottsága.

A kiadásért felelős a BJMT főtitkára
Szedte és tördelte Éliás Mariann

Nyomta a Nagy és Társa Kft., Budapest
Felelős vezető: Fódi Gábor

Budapest, 2009
Megjelent 18 (A/5) ív terjedelemben
250 példányban
HU ISSN 0133-3399

ÚTMUTATÁS A SZERZŐKNEK

Az Alkalmazott Matematikai Lapok csak magyar nyelvű dolgozatokat közöl. A közlésre szánt dolgozatokat e-mailen az aml@math.elte.hu címre kérjük elküldeni az ábrákat tartalmazó fájlokkal együtt. Előnyben részesülnek a LaTeX-ben elkészített dolgozatok.

A kéziratok szerkezeti felépítésének a következő követelményeket kell kielégíteni. A fejlécnek tartalmaznia kell a dolgozat címét és a szerző teljes nevét. A fejléc után egy, képletet nem tartalmazó, legfeljebb 200 szóból álló kivonatot kell minden esetben megadni. A dolgozatot címmel ellátott szakaszokra kell bontani, és az egyes szakaszokat arab sorszámozással kell ellátni. Az esetleges bevezetésnek mindig az első szakaszt kell alkotnia. Az irodalomjegyzék után, a kézirat befejezéseképpen fel kell tüntetni a szerző teljes nevét és a munkahelye (illetve lakása) pontos címét. A dolgozatban előforduló képleteket szakaszonként újrakezdődően, a képlet előtt két zárójel közé írt kettős számozással kell azonosítani. Természetesen nem szükséges minden képletet számozással ellátni. Az esetleges definíciókat és tételeket (segédtételeket és lemmákat) ugyancsak szakaszonként újrakezdődő, kettős számozással kell ellátni. Kérjük a szerzőket, hogy ezeket, valamint a tételek bizonyítását a szövegben kellő módon emeljék ki. Minden dolgozathoz csatolni kell egy angol, német francia vagy orosz nyelvű, külön oldalra gépelt összefoglalót.

Mind az ábrákat, mind a lábjegyzeteket a dolgozat szakaszokra bontásától független, folytatólagos arab sorszámozással kel ellátni. Az ábrák elhelyezését a dolgozat megfelelő helyén, széljegyzetként feltüntetett, ábraazonosító sorszámokkal kell megadni. A lábjegyzetekre a dolgozaton belül az azonosító sorszám felső indexkénti használatával lehet hivatkozni.

Az irodalmi hivatkozások formája a következő. Minden hivatkozást fel kell sorolni a dolgozat végén található irodalomjegyzékben, a szerzők, illetve a társszerzők esetén az első szerző neve szerint alfabetikus sorrendben úgy, hogy a cirill betűs szerzők nevét a Mathematical Reviews átírási szabályai szerint latin betűsre kell átírni. A folyóiratban megjelent cikkekre [1], a könyvekre [5], a kötetben megjelent dolgozatokra [4], a disszertációkra [3] és a gépi program leírásokra [2] a következő minta szerint kell hivatkozni:

[1] Farkas, J.: *Über die Theorie der einfachen Ungleichungen.* Journal für die reine und angewandte Mathematik 124, (1902) 1–27.

[2] Kéri, G.: „DUALSIMP", rutin a CDC 3300-ás gépekre (Magyar Tudományos Akadémia Számítástechnikai és Automatizálási Kutató Intézete, CDC 3300 felhasználói ismertetők 2. 1973. május) 19–20.

[3] Prékopa, A.: *„Sztochasztikus rendszerek optimalizálási problémáiról"*, doktori értekezés. Magyar Tudományos Akadémia, Budapest, 1970.

[4] Prabhu, N. U.: *„Recent research on the ruin problem of collective risk theory"*, in: Inventory Control and Water Storage. Ed. A. Prékopa (János Bolyai Mathematical Society and North-Holland Publishing Company, Amsterdam–London, (1973) 221–228.

[5] Zoutendijk, G.: *Methods of Feasible Directions* (Elsevier Publishing Company, Amsterdam and New York, 1960).

A dolgozatok szövegében az irodalmi hivatkozás számait szögletes zárójelben kell megadni, mint például [5] vagy [4, 76–78]. A szerzők a dolgozatukról 50 darab ingyenes különlenyomatot kapnak. A dolgozatok után szerzői díjat az Alkalmazott Matematikai Lapok nem fizet.

TARTALOMJEGYZÉK

INDEX